Heloisa Maria de Oliveira

Redes de filas com escolha de servidor

CAMPINAS

2012

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

Heloisa Maria de Oliveira

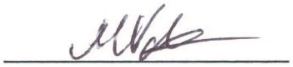
Redes de filas com escolha de servidor

Orientadora: Prof(a). Dr(a). Marina Vachkovskaia

Tese de doutorado apresentada ao Instituto de Matemática, Estatística e Computação Científica da Unicamp para obtenção do título de Doutora em Estatística.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELA ALUNA HELOISA MARIA DE OLIVEIRA E ORIENTADA PELA PROF(A). DR(A). MARINA VACHKOVSKAIA

Assinatura da Orientadora



CAMPINAS
2012

FICHA CATALOGRÁFICA ELABORADA POR
MARIA FABIANA BEZERRA MULLER - CRB8/6162
BIBLIOTECA DO INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E
COMPUTAÇÃO CIENTÍFICA - UNICAMP

Oliveira, Heloisa Maria, 1982-
OL4r Redes de filas com escolha de servidor / Heloisa Maria de Oliveira. – Campinas, SP : [s.n.], 2012.

Orientador: Marina Vachkovskaia.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Redes de computadores. 2. Teoria das filas. 3. Estabilidade. 4. Liapunov, Funções de. 5. Markov, Cadeias de. I. Vachkovskaia, Marina,1975-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em inglês: Queueing network with server choice
Palavras-chave em inglês:
Computer networks
Queueing theory
Stability
Lyapunov function
Markov chain
Área de concentração: Estatística
Titulação: Doutora em Estatística
Banca examinadora:
Marina Vachkovskaia [Orientador]
Christophe Frederic Gallesco
Élcio Lebensztayn
Fábio Prates Machado
Augusto Quadros Teixeira
Data de defesa: 23-11-2012
Programa de Pós-Graduação: Estatística

Tese de Doutorado defendida em 23 de novembro de 2012 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). MARINA VACHKOVSKAIA

Prof(a). Dr(a). CHRISTOPHE FREDERIC GALLESCO

Prof(a). Dr(a). ÉLCIO LEBENSZTAYN

Prof(a). Dr(a). FÁBIO PRATES MACHADO

Prof(a). Dr(a). AUGUSTO QUADROS TEIXEIRA

Dedico este trabalho a todos que de alguma forma contribuíram com o meu progresso como aluna e como Ser.

Agradecimentos

Agradeço,

aos meus pais e irmãos pelo apoio e incentivo durante toda esta jornada da minha vida.

a minha orientadora Profa. Dra. Marina Vachkovskaia pelo apoio fundamental durante todo o programa de doutorado, pelas discussões e ensinamentos enriquecedores que guardarei por toda a minha vida, e principalmente por ser uma excelente pesquisadora, professora, amiga e ser humano.

a todos os professores que compartilharam comigo os seus valorosos ensinamentos durante todo o programa de graduação e de doutorado.

aos membros da banca de defesa e da banca examinadora pelos comentários, sugestões e contribuições que ajudaram a melhorar a qualidade final deste trabalho.

ao meus amigos e funcionários do IMECC pelo apoio, amizade e pelas inúmeras discussões de estatística.

ao $\mathrm{CNP_q}$ (Processo 143187/2008-2) pelo apoio financeiro concedido durante todo o período de doutorado.

à Deus e aos Seres de Luz pelo Amor, pelas pessoas que surgiram na minha vida e que me passaram muitos conhecimentos e ensinamentos, e pelas muitas mudanças que virão para ampliar não só os meus conhecimentos e a minha evolução como ser humano, mas a de todos.

Minha eterna gratidão a todos.

Heloisa Maria Oliveira

“O que eu faço, é uma gota no meio de um oceano.

Mas sem ela, o oceano será menor.”

Madre Teresa de Calcutá

Resumo

Considere as redes de filas com N servidores e $\mathcal{K}$ tipos de trabalho. Suponha que os tipos de trabalho i chegam independentemente à rede de acordo com o processo de Poisson com taxa $\lambda_i > 0$, para todo $i = 1, \ldots, \mathcal{K}$. Para cada trabalho i que chega à rede, um subconjunto fixo e não vazio de servidores (ou filas) é apresentado e, em seguida, de acordo com alguma política previamente estabelecida, os trabalhos são encaminhados para um dos servidores deste subconjunto para serem atendidos. Os tempos dos atendimentos para cada um dos tipos de trabalho i são distribuídos exponencialmente com taxas μ_{ij} dependendo do servidor j escolhido.

Por causa da grande dificuldade na análise deste processo é construído um processo X que tem comportamento semelhante. Neste processo X, os trabalhos i chegam independentemente à rede de acordo com o processo de Poisson com taxa $\lambda_i > 0$, $i = 1, \ldots, \mathcal{K}$, e, em seguida, eles são encaminhados a um servidor j com probabilidade $\pi(\mathbf{x}, i, j)$ para todo estado

$$\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_N) \in \mathbb{R}_+^N$$

que informa a carga horária média total de trabalho de cada um dos N servidores da rede.

Desta forma, para cada trabalho i que chega à rede e que exige um tempo médio $(1/\mu_{ij})$ para ser processado pelos servidores $j = 1, \ldots, N$, adicionará este tempo médio de atendimento à carga horária média total de trabalho do servidor j escolhido.

Os servidores do processo X processam em média "uma parte unitária" da carga horária média total de trabalho por unidade de tempo que está na fila, resultando depois de passar um tempo médio de $(1/\mu_{ij})$, o processamento completo, em média, de um trabalho que exige uma carga horária média de trabalho de $(1/\mu_{ij})$ unidades de tempo.

O objetivo principal deste trabalho é encontrar uma política ideal de escolha de servidor que garanta a estabilidade para estes modelos de rede, cujas taxas dos atendimentos dos servidores alteram-se de acordo com os tipos de trabalho que eles podem processar, utilizando o processo X.

Para o desenvolvimento desta análise foram utilizadas as taxas dos encaminhamentos dos trabalhos e os agrupamentos de servidores apresentados por M.V. Menshikov, I. MacPhee e M. Vachkovskaia [15].

Palavras-chave: Redes de filas, estabilidade, funções de Lyapunov, cadeias de Markov.

Abstract

Consider a queueing network with N servers and $\mathcal{K}$ types of jobs (or customer classes). The types of jobs i, $i = 1, \ldots, \mathcal{K}$, arrive at the system accordingly to independent Poisson process with rate $\lambda_i = 1$, $i = 1, \ldots, \mathcal{K}$. For each arriving customer i a fixed subset of servers (or queues) is presented, and the customer is routed to a server accordingly to some routing policy. Assume that service times for each of the types of jobs i are exponentially distributed with rates μ_{ij} which depend upon the chosen server j.

Because of the great difficulty of the analysis of this process, it was built a process X, which has similar behaviour. In this process X a work i independently arrive at the network according to the Poisson process with rate $\lambda_i > 0$, $i = 1, \ldots, \mathcal{K}$, and then they are routed to a server j with probability $\pi(\mathbf{x}, i, j)$ for every state

$$\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_N) \in \mathbb{R}_+^N$$

which reports the average total number of hours of work of each of the N servers on the network.

Thus, for each job i that comes to the network and that requires an average $(1/\mu_{ij})$ to be processed by servers $j = 1, \ldots, N$, will be added this to the total average hours of work server j chosen.

The servers of the process X execute an "unitary part", on average, of the workload average total work per unit time that is in the queue, resulting after spending an average of $(1/\mu_{ij})$, the executing, on average, of a job that requires a workload average of $(1/\mu_{ij})$ time units.

The main goal of this work is to find an optimal policy choice of server that guarantees stability for these network models, where the service rates of servers change in accordance with the types of jobs they can execute, using the process X.

To develop this analysis it was used the drift rates and clustering presented by M.V. Menshikov, I. MacPhee and M. Vachkovskaia [15].

Key-words: Queueing network, stability, Lyapunov function, Markov chain.

Sumário

Capítulo 1

Introdução

Fila é uma sequência de pessoas ou coisas que se põem uma atrás das outras pela ordem cronológica de chegada. Por exemplo, as filas de espera por um atendimento nos supermercados e bancos, ou as filas de carros que esperam pelo sinal verde do semáforo para seguirem os seus destinos.

Para a maioria das pessoas, as filas representam situações desagradáveis por causa do longo tempo de espera. Por isso, muitas das vezes, as filas são criticadas e enfrentadas com mau humor pelos seus usuários.

A teoria de redes de filas é uma área da pesquisa operacional que utiliza os conceitos sofisticados de processos estocásticos e da matemática aplicada para a compreensão da formação e do desenvolvimento das filas. Esta teoria ajuda analisar a estabilidade de uma rede, possibilitando uma análise mais objetiva e precisa das possíveis situações que podem contribuir para estabilizá-la, ou pelo menos, amenizar o seu sobrecarregamento.

Uma ampla literatura sobre a estabilidade de rede de filas tem aparecido nestas últimas três décadas, nas quais podemos citar os trabalhos de M.V. Menshikov, I. MacPhee e M. Vachkovskaia (2012); A.L. Stolyar e E. Yudovina (2011); A.L. Stolyar e T. Tezcan(2011); M. Bramson (1998 e 2010); J.G. Dai, J.J. Hasenbain e B. Kim (2007); M.J. Luczak e C. McDiarmid (2006); S. Andradóttir e H. Ayhan (2003); S. Asmussen (2003); R.D. Foley e D.R. McDonald (2001); S. Foss e N. Chernova (1998); e R.R. Weber (1978).

Entretanto, a maioria destes trabalhos publicados não analisou a estabilidade das redes de filas cujas taxas dos atendimentos dos servidores alteram-se de acordo

com os tipos de trabalho (ou de cliente) que eles podem atender, principalmente em redes com mais do que dois servidores.

Foss e Chernova (1998) analisaram a estabilidade para este tipo de rede de filas com apenas 2 servidores e comentaram a dificuldade desta análise em redes com mais do que 2 servidores.

Recentemente, Stolyar e Yudovina (2011) apresentaram uma análise para as redes de filas com estas características de dependência entre as taxas dos atendimentos dos servidores e os tipos de trabalho que eles podem atender em redes com mais do que 2 servidores na sua estrutura. Porém, devido à complexidade da análise, Stolyar e Yudovina decidem analisar a estabilidade desta rede fixando as taxas dos servidores, sem levar em consideração o tipo de trabalho (ou de cliente) que chegou ao servidor, chamada por eles de estabilidade global. E, em seguida, eles analisaram a rede de acordo com as taxas dos servidores fixas para cada um dos tipos de trabalho, não levando em consideração o servidor escolhido, chamada de estabilidade local.

Este trabalho considera as redes de filas com N servidores e K tipos de trabalho, e assume os tipos de trabalho i chegando independentemente à rede de acordo com o processo de Poisson com taxa $\lambda_i > 0$, para todo $i = 1, \ldots, K$.

Para cada trabalho i que chega à rede um subconjunto fixo e não vazio de servidores (ou filas) é apresentado e, em seguida, de acordo com alguma política previamente estabelecida, os trabalhos são encaminhados para um dos servidores deste subconjunto para serem processados.

Os tempos dos atendimentos para cada um dos tipos de trabalho i são distribuídos exponencialmente com taxas μ_{ij} dependendo do servidor j escolhido.

O modelo de rede de filas descrito acima é definido como processo Y e analisa a quantidade (ou o número) de trabalhos nos servidores da rede. Perceba que para realizar a análise deste processo é necessário conhecer exatamente todas as taxas dos tipos de trabalho que estão na fila dos servidores e da ordem que chegaram, pois cada trabalho apresenta uma taxa específica de processamento de acordo com o servidor escolhido. Então, por causa da aleatoriedade das taxas dos atendimentos dos tipos de trabalho que estão na fila dos servidores e da ordem que chegaram, esta análise torna-se muito difícil de ser realizada.

Assim, por causa desta dificuldade do conhecimento das taxas dos atendimentos nos servidores é construído um processo X que oferece uma ideia do comportamento deste processo Y.

No processo X, os trabalhos i chegam independentemente à rede de acordo com o processo de Poisson com taxa $\lambda_i > 0$, $i = 1, \ldots, \mathcal{K}$, e, em seguida, eles são encaminhados a um servidor j com probabilidade $\pi(x, i, j)$ para todo estado

$$x = (x_1, \ldots, x_j, \ldots, x_N) \in \mathbb{R}_+^N$$

que informa a carga horária média total de trabalho de cada um dos N servidores da rede.

Desta forma, para cada trabalho i que chega à rede e que exige um tempo médio $(1/\mu_{ij})$ para ser processado pelos servidores $j = 1, \ldots, N$, adicionará este tempo médio para ser processado à carga horária média total de trabalho do servidor j escolhido. E os servidores processam em média "uma parte unitária" da carga horária média total de trabalho por unidade de tempo que está na fila, resultando depois de passar um tempo médio de $(1/\mu_{ij})$, o processamento completo, em média, de um trabalho que exige uma carga horária média de trabalho de $(1/\mu_{ij})$ unidades de tempo.

Os objetivos deste trabalho são

- analisar a estabilidade do processo X;
- verificar que os resultados obtidos pelo processo X oferecem uma boa ideia sobre o comportamento do processo Y com o auxílio das simulações de dois exemplos de rede de filas; e
- avaliar qual ou quais as políticas de escolha de servidor podem oferecer um desempenho eficiente para as redes cujas taxas dos atendimentos dos servidores alteram-se de acordo com os tipos de trabalho que eles podem processar.

O Capítulo 2 apresenta em detalhes a descrição do processo Y e do processo X e, em seguida, introduz as políticas de escolha de servidor analisadas e exibe os resultados elaborados a partir do processo X.

O Capítulo 3 descreve as ideias principais dos algoritmos dos encaminhamentos dos trabalhos aos servidores da rede para cada uma das políticas de escolha

de servidor analisada e, em seguida, apresenta os resultados e as conclusões das simulações. Estes resultados simulados complementam os resultados teóricos fornecidos pelo Capítulo 2 e indicam que o processo X apresenta realmente um comportamento semelhante ao comportamento apresentado pelo processo Y.

Por sua vez, no Capítulo 4 há uma discussão sucinta de como procurar uma política ideal de escolha de servidor para estas redes de filas que apresentam as taxas dos atendimentos dependentes do tipo de trabalho e do servidor escolhido.

Antes de iniciar o Capítulo 2 são apresentados alguns conceitos básicos sobre as cadeias de Markov na Seção 1.1 neste capítulo introdutório. Além disso, há uma breve discussão sobre as redes e os fluxos de redes na Seção 1.2, e por fim, alguns conceitos fundamentais sobre a teoria de filas na Seção 1.3.

1.1 Cadeias de Markov

Definição 1. *Um processo estocástico é uma coleção de variáveis aleatórias denotadas por* $\{X(t),\, t \in T\}$ *indexadas por um parâmetro* t *que usualmente representa o tempo e que estão definidas num mesmo espaço de probabilidade.* □

O conjunto de índices T é o tempo do processo estocástico e $X(t)$ refere-se ao estado do processo no instante t. Dependendo da natureza do tempo, o processo pode ser classificado como sendo um processo a tempo contínuo ou a tempo discreto, ou seja,

- se T é uma sequência contável, por exemplo,

$$T = \{\ldots, -2, -1, 0, 1, 2, \ldots\} \text{ ou } T = \{0, 1, 2, \ldots\},$$

  então, o processo $\{X(t), t \in T\}$ é um processo a tempo discreto, e

- se T é um intervalo, por exemplo,

$$T = \{t : -\infty < t < +\infty\} \text{ ou } T = \{t : 0 < t < +\infty\},$$

  então, o processo $\{X(t),\, t \in T\}$ é definido como processo a tempo contínuo.

Definição 2. *Uma coleção de variáveis aleatórias* $X(t)$*, num espaço de estados* Ω *finito ou infinito enumerável, é uma* cadeia de Markov a tempo discreto *se as variáveis aleatórias satisfazem a condição de Markov, ou seja, se para todo* t*,* $x_0, x_1, \ldots, x$*, tal que* $\mathbb{P}\{X(0) = x_0, X(1) = x_1, \ldots, X(t-1) = x_{t-1}, X(t) = x\} > 0$ *e* $y \in \Omega$*, tem-se*

$$\begin{aligned}
&\mathbb{P}\{X(t+1) = y \mid X(0) = x_0, X(1) = x_1, \ldots, X(t-1) = x_{t-1}, X(t) = x\} \\
&\quad = \mathbb{P}\{X(t+1) = y \mid X(t) = x\} \\
&\quad = p_{xy}
\end{aligned}$$

□

Para as cadeias de Markov, a distribuição condicional de qualquer estado futuro $X(t+1)$ dado o estado presente $X(t)$ e os estados passados $\{X(0), X(1), \ldots, X(t-1)\}$, dependem somente do estado presente e é condicionalmente independente dos estados passados.

Note que p_{xy} repesenta a probabilidade que o processo partindo do estado x estará no estado y. Além disso, usualmente, chama-se o estado inicial de X_0.

Definição 3. *Uma coleção de variáveis aleatórias* $\{X(t), t \geq 0\}$ *é uma* cadeia de Markov a tempo contínuo *se para todo* $s, t \geq 0$ *e* $x,\ y,\ x_u \in \Omega,\ 0 \leq u < s$*, satisfazem a propriedade de Markov, ou seja,*

$$\begin{aligned}
&\mathbb{P}\{X(t+s) = y \mid X(s) = x,\, X(u) = x_u,\, 0 \leq u < s\} \\
&\quad = \mathbb{P}\{X(t+s) = y \mid X(s) = x\}
\end{aligned}$$

□

A cadeia de Markov a tempo contínuo é um processo estocástico que satisfaz a propriedade de Markov, ou seja, que a distribuição condicional do futuro $X(t+s)$ dado o estado presente $X(s)$ e os estados passados $X(u)$, $0 \leq u < s$, depende somente do estado presente e é condicionalmente independente dos estados passados.

Definição 4. *Uma* matriz de transição $\mathbf{P}(x, y) = (p_{xy})_{x,y \in S}$ *é uma matriz* $|\Omega| \times |\Omega|$ *de probabilidades de transições, de tal forma que*

$$p_{xy} \geq 0,\ \ x, y \geq 0;\ \sum_{y=0}^{\infty} p_{xy} = 1,\ \ x = 0, 1, \ldots$$

□

Uma matriz verificando esta propriedade denomina-se de matriz estocástica. A distribuição a n-passos é definida por

$$\mathbf{P}(x,y)^n = \begin{cases} \mathbf{P}(x,y), & n=1\,; \\ \sum\limits_{z\in\Omega} \mathbf{P}(x,z)\mathbf{P}^{n-1}(z,y)\,, & n>1\,. \end{cases}$$

Definição 5. *Uma distribuição é estacionária se ela for invariante com relação à matriz de transição, ou seja,*

$$\text{para todo } y \in \Omega, \quad \pi(y) = \sum_{x\in\Omega} \pi(x)\mathbf{P}(x,y).$$

□

Definição 6. *O tempo para um processo atingir um conjunto (ou estado) B é, por definição,*

$$\tau_B \stackrel{def}{=} \inf\{t \in T : X(t) \in B\}$$

onde T pode ser $\mathbb{N}$ ou $\mathbb{N}^*$. □

Definição 7. *O tempo de retorno de um processo a tempo discreto ao estado $x \in \Omega$ é*

$$\tau_x \stackrel{def}{=} \min\{t > 0 : X(t) = x\}.$$

□

Definição 8. *Considerando f_i como a probabilidade de que iniciando no estado i a cadeia de Markov $X(t)$ retorne a este estado, ou seja,*

$$f_i = \mathbb{P}(\exists t > 0 : X(t) = i \mid X(0) = i).$$

Se $f_i = 1$ o estado é recorrente, já se $f_i < 1$ o estado será transiente. □

Definição 9. *Um estado recorrente $x \in \Omega$ é*

- *recorrente positivo se o tempo médio de retorno ao estado x é finito, ou seja,*

$$\mathbf{E}(\tau_x) < \infty\,, \ e$$

- *recorrente nulo* *se se o tempo médio de retorno ao estado* x *é infinito, ou seja,*

$$\mathbf{E}(\tau_x) = \infty\,.$$

□

Definição 10. *O período de um estado* x *é definido como*

$$\tau_x^* \stackrel{\text{def}}{=} \text{mdc}\{k \geq 1 : \mathbb{P}\{X(k) = x \mid X(0) = x\} > 0\}\,.$$

□

Definição 11. *Um estado* x *é classificado como aperiódico (ou acíclico) se* $\tau_x^* = 1$ *e um estado será periódico (ou cíclico) se* $\tau_x^* = 2$. □

Definição 12. *Um estado é ergódico se for aperiódico e recorrente positivo.* □

Definição 13. *Uma cadeia é ergódica se todos os seus estados são ergódicos.* □

1.2 Redes e fluxos

Este texto segue Rockafellar (1984, Capítulo 1-3) na discussão dos conceitos fundamentais sobre redes e fluxos. Estes conceitos ajudam a compreender melhor as ideias que serão desenvolvidas na construção dos agrupamentos de servidores do próximo capítulo.

O termo redes é amplamente utilizado em inúmeras situações atuais, nas quais podemos citar, as redes de computadores, redes de serviços, redes telefônicas, entre outras.

Definição 14. *Uma rede* $\mathcal{G}$ *é formada por uma tripla* (N, A, F)*, onde* N *é um conjunto finito de vértices;* A *é um conjunto finito de elos ou arestas; e* F *é uma função que atribui a cada elo* $j \in A$ *um par* $(i,\ i') \in N \times N$*, tal que* $i \neq i'$. □

Definição 15. *Uma rede é orientada (ou dirigida) se todos os seus elos são orientados, isto é, cada um dos elos tem um sentido que lhe será associado. Nesta situação, por exemplo, o elo* $j_k \sim (i_{i-k}, i_k)$ *que tem como ponto inicial* i_{i-k} *e como ponto final* i_k*. (O símbolo* $\sim$ *pode ser lido como "corresponde a".)* □

Definição 16. *Um* caminho *(ou* direcionamento*)* P *é uma sequência finita*

$$i_0,\ j_1,\ i_1,\ j_2,\ i_2,\ \ldots,\ j_r,\ i_r\,(r>0)$$

onde cada i_k *é um vértice e* j_k *é um elo, apresentando* $j_k \sim (i_{i-k}, i_k)$ *ou* $j_k \sim (i_k, i_{k-1})$. □

Portanto, um caminho P é uma trajetória que sai de um vértice inicial i_0 até um vértice final i_r, cuja representação é $P : i_0 \to i_r$.

Se $i_0 = i_r$, então o caminho P é um circuito, ou seja, não é possível distinguir o vértice final do vértice inicial. Da mesma forma é possível representar um caminho que sai de um conjunto de vértices S e que chega no conjunto de vértices S', ou seja, $P : S \to S'$.

Definição 17. *Uma* rede *(ou grafo) é* conexa *se existe um caminho que une cada par de vértices distintos da rede.* □

Considerando uma rede com N vértices e assumindo N^+ e N^- dois subconjuntos disjuntos não nulos dos N vértices, de tal forma que N^+ apresenta o subconjunto de vértices cujos fluxos saem e N^- o subconjunto de vértices cujos os fluxos entram, é possível definir um corte em uma rede.

Corte em redes Para compreender o que é um corte numa rede, considere dois conjuntos com vértices arbitrários S e S', e defina o conjunto dos elos da seguinte maneira,

$$[S,\ S']^+ = \{j \in A : j \sim (i,\ i') \text{ com } i \in S \text{ e } i' \in S'\},$$
$$[S,\ S']^- = \{j \in A : j \sim (\ i', i) \text{ com } i \in S \text{ e } i' \in S'\}.$$

Um conjunto com sinal é um conjunto dividido em dois subgrupos, um projetado positivamente e outro negativamente. Um corte de uma rede é um conjunto com sinal Q (com partes positivas e negativas denotadas respectivamente por Q^+ e Q^-), de tal forma que para algum vértice do conjunto S (pode não ser único) existe um Q^+ e Q^- (neste caso, utiliza-se a notação $Q = [S, N \setminus S]$).

Desta forma, é possível escrever um conjunto $Q = [S, N \setminus S]$ como sendo um conjunto formado por um subconjunto $Q^+ = [S, N \setminus S]^+$ de vértices cujos elos saem

de S para S' e por um subconjunto $Q^- = [S, N \setminus S]^-$ de vértices cujos elos saem de S' para S. Portanto, qualquer conjunto definido por $Q = [S, N \setminus S]$ que pode ser particionado em subconjuntos de vértices Q^+ e Q^-.

Definição 18. *Um* corte de uma rede *é um conjunto* Q *formado por conjunto positivo* $Q^+ = [S, N \setminus S]^+$ *e por um conjunto negativo* $Q^- = [S, N \setminus S]^-$. *De tal forma que existe a possibilidade da separação dos conjuntos disjuntos não-nulos de vértices* N^+ *e* N^-. □

Geralmente, associa-se a cada elo $j \in A$ de uma rede um intervalo de capacidade $C(j) = [c^-(j), c^+(j)]$, com $c^-(j) \leq c^+(j)$, onde $c^-(j)$ é a menor capacidade do elo j e $c^+(j)$ é a maior capacidade do elo j.

Definição 19. *Um* fluxo x *é denominado de* viável *com respeito a capacidade se* $x(j) \in C(j)$ *para todo* $j \in A$. □

A *capacidade de um corte* Q é dada por

$$c^+(Q) = \sum_{j \in Q^+} c^+(j) - \sum_{j \in Q^-} c^-(j) \text{ e}$$

$$c^-(Q) = \sum_{j \in Q^+} c^-(j) - \sum_{j \in Q^-} c^+(j).$$

.

Definição 20. *O* problema do fluxo máximo *é maximizar o fluxo que sai do conjunto de vértices* N^+ *para o conjunto de vértices* N^-, *respeitando a capacidade máxima que cada elo pode suportar nos caminhamentos.* □

Definição 21. *O* corte mínimo de uma rede *é um corte que minimiza a capacidade máxima* $c^+(Q)$ *sobre todos os cortes possíveis da rede separando* N^+ *de* N^-. □

Teorema 1 (Teorema do corte mínimo e fluxo máximo)**.** *A solução do problema do fluxo máximo é igual a solução do corte mínimo.*

Demonstração. A demonstração deste teorema pode ser verificada na Seção 3G de Rockafellar (1984, Capítulo 3). □

1.3 Teoria de filas

A teoria de redes de filas foi desenvolvida com o objetivo principal de promover análises matemáticas precisas com o propósito de melhorar a estabilidade das redes. Principalmente, quando a procura por um atendimento é muito maior do que a capacidade da rede, ocasionando o sobrecarregamento dos servidores e inviabilizando a utilização da rede.

Definição 22. *Um processo será estável se este processo for recorrente positivo.* □

Definição 23. *Uma rede de filas é um sistema que consiste em clientes (trabalhos, informações ou materiais) chegando em um servidor durante um determinado período de tempo, esperando para serem atendidos (ou processados), e deixando a rede depois de serem atendidos.* □

As redes de filas são definidas segundo as seguintes características.

C 1. *Padrões de chegadas*

Os modelos de redes de filas mais comuns utilizam as chegadas dos clientes de acordo com os processos de Poisson independentes. Este é um dos processos de contagem mais importante, portanto, os principais conceitos sobre o processo de Poisson são apresentados a seguir.

Definição 24. *Um processo* $\{N(t), t \geq 0\}$ *é chamado de processo de contagem se* $N(t)$ *representando o número total de eventos que podem ocorrer no tempo* t*, satisfaz as seguintes propriedades.*

(I) $N(t) \in \{0, 1, 2, \ldots\}$*;*

(II) Se $s < t$*, então* $N(s) \leq N(t)$*; e*

(III) Para $s < t$*, então* $N(s) - N(t)$ *é o número número de eventos que ocorreram no intervalo* $(s, t]$*.* □

Definição 25. *O processo de contagem* $\{N(t), t \geq 0\}$ *é um processo de Poisson com taxa* $\lambda > 0$*, se*

(I) $N(0) = 0$*;*

(II) $N(t)$ *apresenta incrementos independentes; e*

(III) *o número de eventos em qualquer intervalo* $(s, t]$ *tem distribuição de Poisson com média* λt*, ou seja, para todo* $s, t \geq 0$

$$\mathbb{P}(N(t+s) - N(s) = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad n=0,1,2,\ldots$$

□

Um processo de contagem possui incrementos independentes se os números de eventos que ocorrem em intervalos de tempo disjuntos são independentes. E um processo de contagem tem incrementos estacionários se a distribuição do número de eventos que ocorrem num intervalo de tempo qualquer depende apenas do comprimento do intervalo de tempo.

Segundo Ross (2007, Capítulo 5) uma outra definição sobre o processo de Poisson equivalente à Definição 25 é a seguinte.

Definição 26. *O processo de contagem* $\{N(t), t \geq 0\}$ *é um processo de Poisson a com taxa* $\lambda > 0$*, se*

(I) $N(0) = 0$*;*

(II) *o processo tem incrementos independentes e estacionários; e*

(III) $\mathbb{P}(N(h) = 1) = \lambda h + o(h) \quad e \quad \mathbb{P}(N(h) \geq 2) = o(h), \quad h \to 0.$

□

Um outro fator importante a ser considerado juntamente com os padrões de chegadas dos clientes é a reação dos clientes no momento em que eles chegam à rede. Uma vez que os clientes podem escolher um servidor para serem atendidos, usando alguma regra de escolha previamente estabelecida por eles.

C 2. *Padrões de serviço*

Em geral, os padrões de serviços (ou atendimentos) são exponencialmente distribuídos com parâmetro μ.

No entanto, estes padrões de serviços podem depender do número de clientes que se encontram na fila esperando pelo serviço. Estas situações, nas quais os clientes

escolhem os servidores de acordo com o número de clientes nas filas, são conhecidas como dependência de estado.

C 3. *Disciplina das filas*

A disciplina das filas define a maneira de como os clientes que estão na fila serão atendidos pelos servidores. As disciplinas mais comuns são

(I) o primeiro que chega é o primeiro a sair do sistema, da língua inglesa first in - first out (FIFO), aplicável em atendimentos bancários, de supermercados e de lojas;

(II) o último a chegar é o primeiro a sair, da língua inglesa last in - first out (LIFO), muito comum em sistemas de controle de estoques, onde o item mais recente é o mais fácil de ser apanhado ou manuseado; e

(III) as duas situações gerais das disciplinas de prioridades,

- o cliente com a maior prioridade é atendido pelo servidor independemente do outro cliente que estava sendo atendido, interrompendo o atendimento do outro cliente para que este seja atendido E somente depois que o cliente prioritário deixa a rede é que se reinicia o atendimento do cliente que foi interrompido. Este tipo de situação é chamada de disciplina prioritária preemptiva, sendo muito comum em casos de emergências em hospitais; e
- os clientes com a maior prioridade vão para o início da fila, porém somente são atendidos depois que o cliente que está sendo atendido pelo servidor deixar a rede. Esta situação é conhecida como disciplina de prioridade não-preemptiva.

C 4. *Capacidade da rede*

Algumas redes apresentam uma limitação física por causa da quantidade de espaço, de tal modo que, se as filas alcançarem um comprimento superior a capacidade do espaço, nenhum novo cliente poderá entrar. Desta forma, somente entrarão clientes na fila quando houver um espaço disponível, ou seja, quando um cliente foi atendido e saiu da fila.

As redes que apresentam alguma restrição de espaço físico são definidas como redes finitas, ou seja, existe um limite finito para o comprimento das filas.

Neste trabalho foram consideradas as redes de filas infinitas, ou seja, redes em que não existe um limite finito para o comprimento das filas.

C 5. *Organização dos servidores e da filas*

A organização dos servidores e das filas pode ser

- fila única de clientes que podem ser atendidos por mais do que um servidor da rede; e
- filas individuais para cada servidor.

C 6. *Estágios de serviços*

Existem redes de filas que podem realizar os atendimentos que

- exigem um único estágio (ou um servidor), como por exemplo o pagamento de uma compra, cujos clientes são atendidos por um único servidor e deixam a rede depois que forem atendidos, ou
- exigem vários estágios (ou vários servidores), como no caso de um exame médico que há a necessidade de vários exames até a conclusão da avaliação médica.

Capítulo 2

Modelo e Resultados

Este capítulo apresenta a descrição do modelo de redes de filas cujas taxas dos atendimentos nos servidores alteram-se de acordo com os tipos de trabalho que eles podem processar, definido pelo processo Y.

No entanto, devido à dificuldade na análise do processo Y por causa da necessidade do conhecimento das taxas dos atendimentos que estão na fila dos servidores e da ordem que chegaram, é construído um processo X que tem comportamento semelhante ao apresentado por este processo. Desta forma, os resultados obtidos a partir do processo X dão uma ideia sobre o comportamento do processo Y.

2.1 Descrição do modelo e notação

A análise do comportamento dos modelos de redes de filas foi realizada de acordo com o modelo de supermercado apresentado pelos trabalhos de M.V. Menshikov, V. Sisko e M. Vachkovskaia (2011); e M.V. Menshikov, I. MacPhee e M. Vachkovskaia (2012). O modelo de supermercado consiste na análise das redes de filas que possuem um número fixo de servidores $C_0 = \{1, \ldots, N\}$ e $\mathcal{K}$ tipos de trabalho.

Considere as redes de filas com N servidores e $\mathcal{K}$ tipos de trabalho. Suponha que os tipos de trabalho i chegam independentemente à rede de acordo com um processo de Poisson com taxa $\lambda_i > 0$, para todo $i = 1, \ldots, \mathcal{K}$.

Para cada trabalho i que chega à rede um subconjunto fixo e não vazio de servidores (ou filas) é apresentado e, em seguida, de acordo com alguma política previamente estabelecida, os trabalhos são encaminhados a um dos servidores deste

subconjunto para serem processados.

Os tempos dos atendimentos para cada um dos tipos de trabalho i são distribuídos exponencialmente com taxas μ_{ij} dependendo do servidor j escolhido.

Os trabalhos na fila dos servidores são processados segundo a disciplina FIFO, da língua inglesa *first in - first out*, ou seja, o primeiro que chega na fila é o primeiro a ser atendido, e os trabalhos somente deixam a rede quando forem completamente processados.

Exemplo 1. *Considere um modelo de rede de filas que possui a capacidade de atender* 3 *tipos de trabalho e que apresenta* 3 *servidores que podem realizar os processamentos dos trabalhos que chegam à rede. No entanto, para cada trabalho* i *que chega à rede existem somente* 2 *servidores que apresentam a capacidade de realizar o processamento com taxas dos atendimentos* μ_{ij}*, para todo* $i = 1, 2, 3$ *e* $j = 1, 2, 3$*, como ilustra a figura abaixo.*

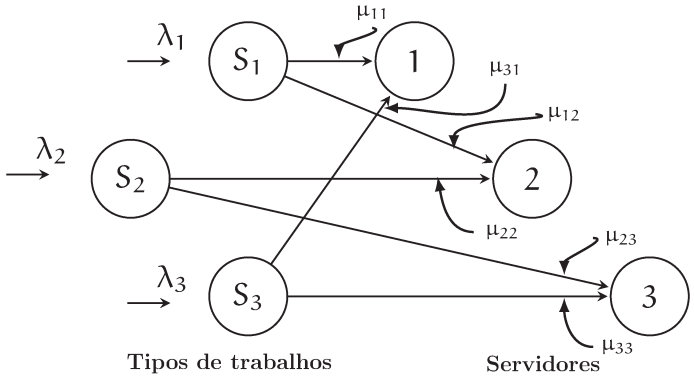


Figura 2.1: Representação do exemplo de rede.

Seja $\mathcal{P}(C_0)$ todos os subconjuntos não vazios e fixos dos N servidores da rede. Para cada um dos subconjuntos de servidores de $\mathcal{P}(C_0)$ existem as chegadas de um tipo específico de trabalho com taxa $\lambda_i > 0$, $i = 1, \ldots, \mathcal{K}$ que os servidores deste subconjunto podem processar. Contudo, perceba que alguns servidores dos subconjuntos eventualmente podem processar mais do que um tipo de trabalho. Desta forma, todos os tipos de trabalho que podem ser processados por um subconjunto $C \subseteq \{1, \ldots, N\}$ de servidores da rede são representados por $\mathcal{N}(C)$.

O modelo de rede de filas descrito acima é definido como um processo de Markov $Y = \{Y(t),\, t \in \mathbb{R}^+\}$ que analisa a quantidade (ou o número) de trabalhos nos

servidores da rede.

Desta forma, para o processo Y, os trabalhos i chegam independentemente à rede de acordo com o processo de Poisson com taxa $\lambda_i > 0$, para todo $i = 1, \ldots, \mathcal{K}$, e, em seguida, eles são encaminhados para um servidor j da rede com probabilidade $\pi(\mathbf{y}, i, j)$ para todo estado

$$\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_N)$$

que apresenta a configuração atual dos trabalhos na fila dos N servidores da rede,

$$\mathbf{y}_j = (\mathcal{Y}_1, \ldots, \mathcal{Y}_l)$$

é o vetor que apresenta a sequência dos tipos de trabalho $\mathcal{Y}_l \in \{1, \ldots, \mathcal{K}\}$ que estão na fila do servidor j na posição l da configuração atual da rede, cuja taxa de atendimento será $\mu_{\mathcal{Y}_l j}$.

Perceba que para realizar a análise deste processo é necessário conhecer exatamente todas as taxas dos tipos de trabalho que estão na fila dos servidores e na ordem que chegaram, pois cada trabalho apresenta uma taxa específica de processamento de acordo com o servidor escolhido. Por causa disso, a matriz $\mathbf{y}$ guarda todas as informações da configuração atual da rede, cujas as colunas da matriz representam um dos servidores da rede e cada linha informa a taxa específica do atendimento do trabalho i que se encontra na fila destes servidores.

Note que devido à aleatoriedade da ordem de chegada das taxas dos atendimentos na fila dos servidores esta análise torna-se muito difícil de ser realizada.

Então, baseando-se no processo de Markov Y, é possível construir um processo de Markov $X = \{X(t),\ t \in \mathbb{R}^+\}$ que analisa a carga horária média total de trabalho que está nos servidores, ao invés de analisar a quantidade (ou o número) de trabalhos que está nos servidores.

No processo X, os trabalhos i chegam independentemente à rede de acordo com a taxa Poisson $\lambda_i > 0$, para todo $i = 1, \ldots, \mathcal{K}$, e, em seguida, eles são encaminhados a um servidor j da rede com probabilidade $\pi(\mathbf{x}, i, j)$ para todo estado

$$\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_N) \in \mathbb{R}_+^N$$

que informa a carga horária média total de trabalho para cada um dos N servidores da rede.

Assim, para cada trabalho i que chega à rede e que exige um tempo médio $(1/\mu_{ij})$ para ser processado pelos servidores $j = 1, \ldots, N$, adicionará este tempo médio de $(1/\mu_{ij})$ à carga horária média total de trabalho do servidor j escolhido, ou seja, $x_j = x_j + (1/\mu_{ij})$.

Os servidores do processo X processam em média "uma parte unitária" da carga horária média total de trabalho por unidade de tempo que está na sua fila, resultando depois de passar um tempo médio de $(1/\mu_{ij})$, o processamento completo, em média, de um trabalho que exige em média uma carga horária de trabalho de $(1/\mu_{ij})$ unidades de tempo.

As eventuais mudanças que podem existir no processo X são analisadas de acordo com os saltos e as taxas dos saltos de acordo com a carga horária média total de trabalho.

Os saltos de uma configuração (ou estado atual) $\mathbf{x}$ da rede são fornecidos pelo vetor δ_j/μ_{ij} ou $-\delta_j$ onde $\delta_j = (0, \ldots, 1, \ldots, 0)$, indicando se houve a entrada de algum trabalho i, de acordo com o aumento da carga horária média total de trabalho do servidor j. E uma saída, se o servidor j processou, em média, um trabalho de carga horária média $(1/\mu_{ij})$, indicada pela j-ésima coordenada do vetor.

Portanto, se um trabalho i chegou à rede e foi encaminhado para o servidor j com taxa $\sum_{i \in \mathcal{N}(C_0)} \pi(\mathbf{x}, i, j)\lambda_i$, consequentemente haverá na j-ésima coordenada do vetor $\mathbf{x}$ um salto de $(1/\mu_{ij})$ que corresponde ao aumento da carga horária média total de trabalho do servidor j, ou seja, $\mathbf{x} \mapsto \mathbf{x} + \delta_j/\mu_{ij}$ ou $\mathbf{x} = (x_1, \ldots, x_j + 1/\mu_{ij}, \ldots, x_N)$, para todo $i = 1, \ldots, \mathcal{K}$ e $j = 1, \ldots, N$. E existirá um salto para baixo na j-ésima coordenada de $\mathbf{x}$ quando o servidor j processar por completo, em média, um trabalho que exige uma carga horária média de trabalho de $(1/\mu_{ij})$, depois de passar um tempo médio $(1/\mu_{ij})$ a partir do início do seu processamento, ou seja, $\mathbf{x} \mapsto \mathbf{x} - \delta_j$ ou $\mathbf{x} = (x_1, \ldots, x_j - 1, \ldots, x_N)$.

Nas situações em que existem $x_j \mapsto x_j - 1 < 0$ é definido $x_j \mapsto 0$, para todo $j \in N$. Isso significa que os servidores só processam os trabalhos que estão na fila e não realizam nenhum trabalho no tempo ocioso.

Este trabalho considerou também o passeio aleatório $\xi(t)$. Para o passeio aleatório $\xi(t)$, se um trabalho i chegou à rede e foi encaminhado ao servidor j com taxa $\sum_{i \in \mathcal{N}(C_0)} \pi(\xi, i, j)\lambda_i$, haverá na j-ésima coordenada do vetor ξ um salto de $(1/\mu_{ij})$ que corresponde ao aumento da carga horária média total de trabalho do servidor j, ou seja, $\xi \mapsto \xi + \delta_j/\mu_{ij}$ ou $\xi = (\xi_1, \ldots, \xi_j + 1/\mu_{ij}, \ldots, \xi_N)$, para todo $i = 1, \ldots, \mathcal{K}$ e $j = 1, \ldots, N$. E existirá um salto para baixo na j-ésima coordenada de ξ quando o servidor j processar por completo, em média, um trabalho que exige uma carga horária média de trabalho de $(1/\mu_{ij})$, depois de passar um tempo médio $(1/\mu_{ij})$ a partir do início do seu processamento, ou seja, $\xi \mapsto \xi - \delta_j$ ou $\xi = (\xi_1, \ldots, \xi_j - 1, \ldots, \xi_N)$.

Note que o passeio aleatório $\xi(t)$ apresenta a mesma descrição que processo X, mas sem a exigência de que seja positivo. O passeio aleatório $\xi(t)$ ajuda a entender as situações nas quais existe uma grande quantidade de pedidos de um determinado tipo de produto e que as empresas precisam liberar, mas devido à indisponibilidade do produto exigido com o tempo estas empresas acabam ficando com os seus pedidos em pendência.

Para facilitar o entendimento sobre o processo X, imagine um supermercado onde os clientes i escolhem um servidor que pode atendê-los. Em seguida, usando alguma política de escolha, eles encaminham-se para a fila do servidor j escolhido com o seu carrinho de compras (ou de trabalho) para ser processado. Os servidores escolhidos demoram em média $(1/\mu_{ij})$ minutos para processar toda a compra do cliente, levando em média um minuto para registrar cada produto comprado no caixa.

Portanto, somente depois de passar um tempo médio de $(1/\mu_{ij})$ minutos a partir do momento que o trabalho começou a ser processado, o servidor conseguirá processar por completo, em média, um trabalho (ou uma compra) que tem uma carga horária média de trabalho de $(1/\mu_{ij})$ minutos.

Note que no processo Y todas as informações da configuração atual da rede são guardadas numa matriz $\mathbf{y}$. Enquanto que no processo X, à medida que os tipos de trabalho são encaminhados aos servidores escolhidos, automaticamente estes trabalhos são transformados em cargas horárias médias de trabalho que, por sua vez, são acumuladas na carga horária média total de trabalho dos servidores

escolhidos.

Desta forma, o processo X não apresenta uma matriz de informações sobre a configuração atual da rede, mas fornece um vetor que apresenta a carga horária média total de trabalho que cada um dos servidores da rede terão de processar.

Portanto, pode-se perceber que o processo X guarda de forma muito mais econômica as informações cruciais sobre o processo Y, apresentando um comportamento semelhante ao obtido pelo processo Y.

Uma outra maneira de entender esta relação entre os processos Y e X pode ser vista pela Tabela 2.1.

Tabela 2.1: Representação do processo Y e do processo X.

	Processo Y	Processo X
Taxas de chegadas	λ_i	λ_i
Taxas de saídas	μ_{ij}	1
Quantidade de trabalho que chega com taxa λ_i	1	$1/\mu_{ij}$
Quantidade de trabalho que sai depois de passar um tempo médio de $(1/\mu_{ij})$ a partir do momento que o trabalho começou a ser processado pelo servidor	1	1

para todo $i \in \mathcal{N}(C)$ e para todo $j \in C = \{1, \ldots, N\}$

Os objetivos deste trabalho são

- analisar a estabilidade do processo X;
- verificar que os resultados obtidos pelo processo X oferecem uma boa ideia sobre o comportamento do processo Y com o auxílio das simulações de dois exemplos de redes de filas; e
- avaliar qual ou quais as políticas de escolha de servidor que podem oferecer um desempenho eficiente para as redes cujas taxas dos atendimentos dos servidores alteram-se de acordo com os tipos de trabalho que eles podem processar.

Para garantir que os servidores não possam ser decompostos em agrupamentos que sejam independentes dos tipos de trabalho que chegam à rede, assume-se que a

rede bipartida G com vértices $\mathcal{N}(C_0) \cup C_0$ e elos $E = \{(S_i, j) : j \in S_i\}$ seja conexa, onde $S_i \subset \{1, \ldots, N\}$ são os servidores que podem processar o tipo de trabalho i, para todo $i = 1, \ldots, \mathcal{K}$.

Descrição das políticas de escolha de servidor

Seja $\Delta_0 = \{\pi \in [0,1]^N : \sum_j \pi(j) = 1\}$ as probabilidades possíveis dos encaminhamentos dos trabalhos aos N servidores da rede. E para cada trabalho i que chega à rede é definido $\Delta_i = \{\pi \in \Delta_0 : \pi(j) = 0 \text{ para } j \notin S_i\}$ como sendo as probabilidades dos encaminhamentos dos trabalhos i aos servidores $j \in S_i$, para todo $i = 1, \ldots, \mathcal{K}$.

Definição 27. *As políticas de escolha de servidor são definidas por*

$$\pi : \mathbb{R}^N \times \{1, 2, \ldots, |\mathcal{N}(C_0)|\} \to \Delta_0 \quad \textit{tal que } \pi(\mathbf{x}, i) \in \Delta_i \,.$$

Isto significa que cada trabalho i que chega à rede quando o estado do processo é $\mathbf{x}$ é encaminhado para a fila do servidor $j \in S_i$ com probabilidade $\pi(\mathbf{x}, i, j)$, cujo espaço de políticas é Π. □

Definição 28. *As políticas estáticas de escolha de servidor são regras ou políticas que encaminham os trabalhos para um servidor $j \in S_i$ com probabilidade $\pi(i, j)$, independentemente da configuração atual $\mathbf{x}$ da rede.* □

O espaço das políticas estáticas é representado por Π^{Estat} e as políticas por $\pi(i)$.

Definição 29. *A política de escolha de servidor cujos trabalhos são encaminhados à fila mais rápida, do inglês join the fastest queue (JFQ), é definida por*

$$\pi(i, j) = \begin{cases} 1/|\mathcal{D}_i(\mathbf{x})| & \textit{se } j \in \mathcal{D}_i(\mathbf{x}) \,, \\ 0 & \textit{caso contrário} \,, \end{cases} \tag{2.1}$$

$$\textit{onde } \mathcal{D}_i(\mathbf{x}) = \{j \in S_i : \mu_{ij} = \max_{l \in S_i} \mu_{il}\} \,.$$

□

Portanto, a política JFQ encaminha os trabalhos i aos servidores que apresentam a taxa de atendimento mais rápida dentro de um conjunto dos servidores que podem realizar o processamento.

É importante ressaltar que os encaminhamentos dos trabalhos para as filas mais rápidas podem não garantir a estabilidade da rede. Esta situação pode acontecer, por exemplo, quando uma rede possui um servidor com a capacidade de processamento muito mais rápida do que os demais servidores para qualquer tipo de trabalho. Desta forma, existirá o sobrecarregamento de algum servidor da rede, inviabilizando o uso da política JFQ.

Portanto, a fim de comparar de modo eficiente as políticas estáticas e dinâmicas, este trabalho considera somente as políticas estáticas que apresentam as probabilidades adequadas dos encaminhamentos dos trabalhos, de tal modo que garantem a estabilidade da rede, denominadas de melhor política estática JSW_b e que são apresentadas na Definição (33). Mais informações e detalhes sobre esta política também serão exibidos no próximo capítulo deste trabalho.

Definição 30. *A política de escolha de servidor cujos trabalhos são encaminhados à fila do servidor* j *que apresenta a menor carga horária média total de trabalho da fila, ponderada pelo tempo médio que o trabalho* i *possa ser processado pelo servidor escolhido, da língua inglesa join the smallest weighted by average service time workload (JSWA), é dada por*

$$\pi(\mathbf{x}, i, j) = \begin{cases} 1/|\mathcal{Q}_i(\mathbf{x})| & \text{se } j \in \mathcal{Q}_i(\mathbf{x}) \,, \\ 0 & \text{caso contrário} \,, \end{cases} \tag{2.2}$$

$$\text{onde } \underline{x}_i = \min_{l \in S_i} (x_l/\mu_{il}) \;\; \text{e} \;\; \mathcal{Q}_i(\mathbf{x}) = \left\{ j \in S_i : \underline{x}_i = x_j/\mu_{ij} \right\}.$$

□

Assim, para a política JSWA, todos os trabalhos i são encaminhados aos servidores que apresentaram a menor carga horária média total de trabalho da fila, ponderada pelo tempo médio de processamento que o servidor j escolhido poderá processar.

Definição 31. *A política de escolha de servidor cujos trabalhos são encaminhados à fila que apresenta a menor carga horária média de trabalho do sistema (tempo médio na fila + tempo médio do processamento do trabalho) ponderada pelo tempo médio que o trabalho* i *possa ser processado pelo servidor* j *escolhido, da língua inglesa join*

the smallest weighted by average service time system workload (JSWAS), é definida por

$$\pi(\mathbf{x}, i, j) = \begin{cases} 1/|\mathcal{B}_i(\mathbf{x})| & \text{se } j \in \mathcal{B}_i(\mathbf{x}), \\ 0 & \text{caso contrário}, \end{cases} \tag{2.3}$$

$$\text{onde } \underline{x}_i^* = \min_{l \in S_i} \left[\left(x_l + 1/\mu_{il} \right) / \mu_{il} \right] \text{ e}$$

$$\mathcal{B}_i(\mathbf{x}) = \left\{ j \in S_i : \underline{x}_i^* = \left(x_j + 1/\mu_{ij} \right) / \mu_{ij} \right\}.$$

□

Portanto, para a política JSWAS todos os trabalhos i são encaminhados aos servidores que apresentaram a menor carga horária média total de trabalho considerando o aumento da carga horária média total de trabalho que o trabalho poderá adicionar aos servidores escolhidos, ponderada pelo tempo médio que os trabalhos possam ser processados pelos servidores j escolhidos.

Motivação para encontrar a política ideal de escolha de servidor

Considerando as redes de filas cujas taxas dos atendimentos são independentes dos tipos de trabalho, ou seja, $\mu_{ij} = \mu_j$, pode-se tentar encontrar a política "ideal" de escolha de servidor através da minimização do tempo médio que um cliente qualquer passa na rede, de acordo com as políticas estáticas Π^{Estat}, dado por

$$W^{\Pi^{Estat}} = \sum_{j=1}^{N} \frac{\sum_{i=1}^{K} \lambda_i \pi(i, j)}{\mu_j - \sum_{l=1}^{K} \lambda_l \pi(l, j)}. \tag{2.4}$$

Desta forma, é possível encontrar o valor mínimo de $W^{\Pi^{Estat}}$ cuja solução satisfaça o sistema de inequações lineares

$$\begin{cases} \sum_{i=1}^{K} \lambda_i \pi(i, j) < \mu_j \quad \text{para } j = 1, \ldots, N, \\ \pi(i, j) \geq 0 \text{ e } \sum_{j \in S_i} \pi(i, j) = 1, \end{cases} \tag{2.5}$$

onde $\pi(i, j)$ é a probabilidade do trabalho i ser encaminhado ao servidor j.

Entretanto, este cálculo é difícil de ser realizado, tornando-se ainda mais complicado quando as políticas dinâmicas de escolha de servidor são consideradas.

Contudo, existe uma outra possibilidade de procurar a política ideal de escolha de servidor que consiste em otimizar a cauda do tempo que um cliente qualquer passa na rede.

Portanto, de acordo com Ross (2007, Seção 8.3), o tempo que um cliente qualquer passa na rede para as políticas estáticas $W^{\Pi^{Estat}}$, tem-se

$$\mathbb{P}(W^{\Pi^{Estat}} \geq t) = \sum_{j=1}^{N} \sum_{i=1}^{\mathcal{K}} \lambda_i \pi(i,j) e^{-\left(\mu_j - \sum_{i=1}^{\mathcal{K}} \lambda_i \pi(i,j)\right)t}.$$

Note que é mais razoável tentar encontrar a solução de (2.5) de tal forma que

$$\min_{j=1,\ldots,N} \left\{ \mu_j - \sum_{i=1}^{\mathcal{K}} \lambda_i \pi(i,j) \right\} \text{ é máximo.} \tag{2.6}$$

Isso acontece porque o termo responsável pelo comportamento da cauda do tempo médio que um cliente qualquer passa muito tempo na rede é fornecido por $\mu_j - \sum_{i=1}^{\mathcal{K}} \lambda_i \pi(i,j)$.

Portanto, se maximizar (2.6), consequentemente é possível otimizar a probabilidade que um cliente qualquer permaneça muito tempo na rede.

Então, a partir desta motivação é possível estabelecer as políticas ideais de escolha de servidor usando a carga horária média total de trabalho que está na fila dos servidores, como descreve o processo X, de tal forma que

$$\min_{j=1,\ldots,N} \left\{ 1 - \sum_{i=1}^{\mathcal{K}} \pi(i,j) \frac{\lambda_i}{\mu_j} \right\} \text{ é o máximo possível.} \tag{2.7}$$

Isto significa que a velocidade da fila mais lenta deve ser a mais rápida possível.

Em virtude desta motivação, assume-se a velocidade da fila do servidor j para qualquer política de escolha π no instante t, para todos os estados de $\mathbf{x}$ com $x_j > 0$, como sendo

$$V(j, \mathbf{x}, \pi) = \sum_{i \in N(C_0)} \pi(\mathbf{x}, i, j) \frac{\lambda_i}{\mu_{ij}} - 1. \tag{2.8}$$

Note que para qualquer política estática de escolha de servidor, cuja política não depende da configuração atual $\mathbf{x}$ da rede, a representação dada acima será

$$V(j,\pi) = \sum_{i \in N(C_0)} \pi(i,j)\frac{\lambda_i}{\mu_{ij}} - 1. \tag{2.9}$$

Definição 32. *Um agrupamento é um subconjunto não vazio de servidores* $C \subseteq C_0$.

□

Assim, para qualquer agrupamento C de servidores e para qualquer subconjunto de políticas estáticas $\Pi' \subset \Pi^{\text{Estat}}$, defina

$$\mathcal{V}(C;\mu_{ij},\Pi') = \min_{\pi\in\Pi'} \max_{j\in C} V(j;\pi)\ , \tag{2.10}$$

ou seja, o mínimo (sobre todas as políticas Π') da velocidade máxima da fila dos servidores que estão dentro do agrupamento C.

A política ideal de escolha de servidor é obtida através da análise da velocidade dos servidores (ou das filas) dentro dos agrupamentos de servidores submetidos a um conjunto de políticas previamente estabelecidas.

Definição 33. *A melhor política estática JFQ_b (política ideal de escolha de servidor) é aquela política que oferece*

- *a velocidade da fila mais lenta é a mais rápida possível,*
- *o número de filas que crescem com velocidade máxima é o menor possível, e*
- *as velocidades das filas são decrescentes.*

□

2.2 Resultados

O objetivo principal deste trabalho é encontrar as políticas de escolha de servidor que estabilizem a rede de filas, ou seja, as políticas que são capazes de manter as velocidades (2.8) as menores possíveis de tal maneira que

- a velocidade da fila mais lenta é a mais rápida possível;
- o número de filas que crescem com velocidade máxima é o menor possível, e

- as velocidades das filas são decrescentes.

Para isso, primeiramente é necessário verificar se é possível decompor o conjunto de servidores $C_0 = \{1, \ldots, N\}$ da rede numa hierarquia de agrupamentos disjuntos C_1, ..., C_K de servidores para algum $1 \leq K \leq N$, como apresenta o teorema descrito abaixo.

Teorema 2.2.1. *Considerando* $V_1 = \mathcal{V}(C_0; \mu_{ij}, \Pi^{Estat})$*, pode-se decompor o conjunto de servidores* $C_0 = \{1, \ldots, N\}$ *numa hierarquia de agrupamentos disjuntos* C_1, ..., C_K *de servidores para algum* $1 \leq K \leq N$ *com as seguintes propriedades.*

(I) C_1 *é o único agrupamento de* C_0 *tal que* $\mathcal{V}(C_0; \mu_{ij}, \Pi^{Estat}) = V_1$ *e* $|C_1|$ *é mínimo.*

(II) Se $C_1 \neq C_0$*, então para cada etapa* $k = 2, \ldots,$ *tem-se*

$$V_k = \mathcal{V}(C_0 \setminus \cup_{n=1}^{k-1} C_n \,;\, \mu_{ij}, \Pi_{k-1})$$

onde Π_{k-1} *é o conjunto de políticas estáticas que apresentam a velocidade* V_n *do agrupamento* C_n *para* $n = 1, \ldots, k-1$*. Tendo* $C_k \subseteq C_0 \setminus \cup_{n=1}^{k-1} C_n$ *como sendo o único agrupamento que satisfaz* $\mathcal{V}(C; \mu_{ij}, \Pi_{k-1}) = V_k$ *com valor mínimo de* $|C|$*. De tal forma que para cada etapa* k *da decomposição, obtêm-se* $V_k < V_{k-1}$*, onde* $K \leq N$ *e* $\cup_{n=1}^{K} C_n = C_0$*. Então, a decomposição hierárquica mínima está completa.*

(III) Π_K *é o conjunto não vazio de políticas estáticas. Para cada* $\pi \in \Pi_K$ *e para cada* $j \in C_k$*,* $k = 1, \ldots, K$*, tem-se*

$$V_k = \sum_{i \in \mathcal{N}(C_0)} \pi(i, j) \frac{\lambda_i}{\mu_{ij}} - 1 \,.$$

Isto significa que a velocidade V_k dos servidores do agrupamento C_k informa o comportamento da carga horária média total de trabalho nos servidores segundo as taxas dos atendimentos μ_{ij}, $i \in \mathcal{N}(C_k)$, $j \in C_k$, usando (2.6).

Teorema 2.2.2. *Se* $V_1 > 0$*, então o processo* X *com respeito à carga horária média total de trabalho na fila dos servidores é transiente para qualquer política estática de escolha de servidor. Se* $V_1 < 0$*, então o processo* X *é recorrente positivo (estável) para a melhor política estática de escolha de servidor* JFQ_b*.*

Definição 34. *Um processo* W *será* recorrente em forma *se o processo* $(W_i - W_{i+1})$ *for recorrente para todo* $i = 1, \ldots, n$ *e* $n + 1 := 1$.

Sob o ponto de vista da dinâmica do passeio aleatório $\xi(t)$ submetido às políticas dinâmicas, ou seja, JSWA e JSWAS, pode-se concluir que este passeio é recorrente em forma para cada agrupamento C_k de servidores, $k = 1, \ldots, K$, como estabelece o Teorema 2.2.3.

Teorema 2.2.3. *O passeio aleatório* $\xi(t)$ *segundo as políticas JSWA e JSWAS é recorrente em forma para cada agrupamento* C_k *de servidores,* $k = 1, \ldots, K$.

Portanto, estes resultados demonstram que a velocidade V_k pode ser interpretada como uma taxa de crescimento das filas que estão dentro de um agrupamento C_k de servidores.

Observação 1. *A política ideal de escolha de servidor também pode ser encontrada através da solução de um problema de programação linear.*

2.3 Demonstrações dos resultados

2.3.1 Preliminares para a prova do Teorema 2.2.1

A decomposição hierárquica dos servidores foi baseada na ideia dos fluxos das chegadas dos tipos de trabalho que chegam à rede e no teorema corte mínimo e fluxo máximo de Ford e Fulkerson (1974), apresentados brevemente no Capítulo 1, ou seja, que a decomposição hierárquica mínima dos servidores coincide com a decomposição obtida pelo fluxo máximo.

Para facilitar esta ideia sobre a decomposição dos servidores, uma discussão sucinta sobre o assunto é introduzida à medida que as notações e definições são apresentadas no decorrer da demonstração.

Decomposição dos servidores baseada nos fluxos das chegadas dos trabalhos.

Para cada agrupamento $C \subset C_0$, considere $\mathcal{P}(C)$ como sendo todos os subconjuntos não vazios de servidores que podem ser formados pelo agrupamento C de servidores, $\mathcal{N}(C_0) = \{S \in \mathcal{P}(C_0) : \exists\, i = \{1, \ldots, \mathcal{K}\} \text{ onde } S = S_i\}$ e

$\mathcal{N}(C) = \{S_i \in \mathcal{N}(C_0) : S_i \subseteq C\}$ como sendo todos os tipos de trabalho que chegam à rede com taxa $\lambda_i > 0$ que podem ser processados pelos servidores que estão dentro do agrupamento C.

Definição 35. *O sistema restrito C é um agrupamento de servidores cujas chegadas dos tipos de trabalho $\mathcal{N}(C)$ não podem ser mais decompostos em novos agrupamentos de servidores.* □

O sistema restrito-reduzido D de um agrupamento de servidores é encontrado através da remoção dos servidores $(C_0 \setminus D)$ juntamente com os fluxos das chegadas dos tipos de trabalho $\mathcal{N}(C_0 \setminus D)$ que formam um sistema restrito.

Definição 36. *O sistema restrito-reduzido D é um subconjunto de servidores que pode ser reduzido sucessivamente durante o processo de decomposição hierárquica, ou seja, pode ser reduzido através das sucessivas remoções de servidores que formam um sistema restrito C.* □

Para tornar mais compreensível como funcionam os encaminhamentos dos tipos de trabalho aos servidores de uma rede de acordo com o sistema restrito e de acordo com o sistema restrito-reduzido, considere o seguinte exemplo de rede apresentado pela Figura 2.2.

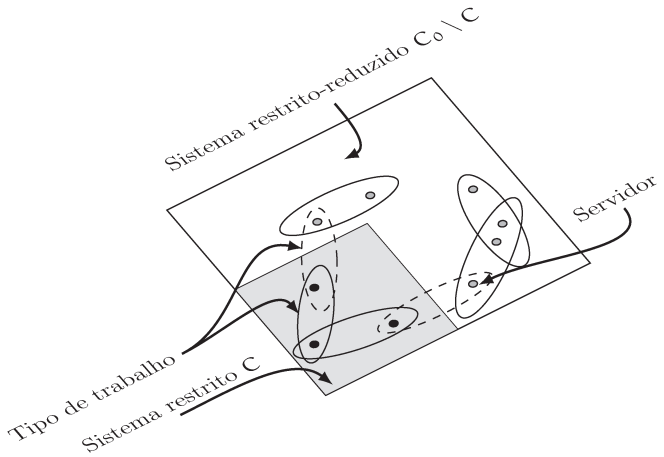


Figura 2.2: Exemplo de sistema restrito e sistema restrito-reduzido

Assim, segundo o exemplo fornecido pela Figura 2.2, é definido que

- ◇ se chegou algum tipo de trabalho que só pode ser processado pelos servidores do sistema restrito, então este trabalho vai ser encaminhado para um dos servidores de cor preta para a realização do seu processamento, segundo alguma política de escolha;
- ◇ se chegou algum tipo de trabalho que pode ser processado tanto pelos servidores do sistema restrito como pelos servidores do sistema reduzido, como mostram as elipses tracejadas, obrigatoriamente estes tipos de trabalho vão ser encami-nhados para os servidores do sistema reduzido, ou seja, para os servidores cinza, cujo servidor será escolhido segundo alguma política previamente estabelecida; e
- ◇ se chegou algum tipo de trabalho que só pode ser processado pelos servidores do sistema reduzido, então este trabalho vai ser encaminhado para um dos servidores de cor cinza para a realização do seu processamento, segundo alguma política de escolha.

Para cada agrupamento $C \subseteq D$ do sistema restrito, existe todos os tipos de trabalho que podem ser processados pelo sistema restrito, que surge a partir de um sistema restrito-reduzido D, isto é,

$$\mathcal{N}_D(C) = \{S \in \mathcal{P}(C) : S_i \subseteq D, \lambda_i(D) > 0\}.$$

Para estabelecer quais os servidores formam um sistema restrito C e um sistema restrito-reduzido D é necessário definir a velocidade média restrita-reduzida de um agrupamento $C \subseteq D$ como apresentada a seguir.

Definição 37. *A velocidade média restrita-reduzida de um agrupamento* $C \subseteq D$ *é dada por*

$$W_D(C) = \frac{1}{|C|} \sum_{j \in C} \left(\sum_{i \in \mathcal{N}_D(C)} \pi(i,j) \frac{\lambda_i(D)}{\mu_{ij}} - 1 \right). \tag{2.11}$$

□

Desta forma, a velocidade do sistema restrito C_k da k-ésima etapa da decomposição é estabelecida pela velocidade $V_k = \max_{C \subseteq D_k} W_{D_k}(C)$, obtida pelo sistema restrito-reduzido $D_k = (C_0 \setminus \cup_{n=1}^{k-1} C_n)$. E os servidores que apresentam esta velocidade V_k formam o sistema restrito C_k.

Observação 2. *Note que os tipos de trabalho $\mathcal{N}(C)$ que chegam à rede não podem ser encaminhados para os servidores que estão fora do agrupamento C. Entretanto, é possível que exista outros tipos de trabalho que podem ser encaminhados para os servidores que estão no agrupamento C.* □

A demonstração do Teorema 2.2.1 considera somente as políticas estáticas Π^{Estat}, ou seja, as políticas cujos encaminhamentos dos trabalhos não consideram a configuração atual da rede no momento em que chegam à rede.

O Teorema 2.2.1 demonstra que é possível decompor eventualmente a rede em sistemas restritos e em sistemas restritos-reduzidos até obter uma hierarquização completa.

Para isso, é necessário definir o conjunto não vazio dos tipos de trabalho $S^D \in \mathcal{P}(D)$ que o sistema restrito-reduzido D pode processar, ou seja,

$$S^D = \{S_i \in \mathcal{N}(C_0) : S_i \cap D \neq \emptyset\}.$$

A união dos fluxos de todos os tipos de trabalho que chegam ao conjunto não vazio dos tipos de trabalho S^D é a taxa total das chegadas do sistema restrito-reduzido D, dada por

$$\lambda_S(D) = \sum_{i \in S^D} \lambda_i \ . \tag{2.12}$$

Combinando esta última informação com a taxa total das chegadas dos trabalhos do sistema restrito-reduzido D, onde $D = C_0$, é possível concluir que existe uma política estática capaz de decompor eventualmente o sistema de rede original com restrição a um agrupamento restrito C_1 e que o sistema restrito-reduzido $D = (C_0 \setminus C_1)$ pode ser decomposto em mais vezes, bastando apenas repetir sequencialmente novamente a decomposição do sistema a partir do sistema restrito-reduzido $D = (C_0 \setminus C_1)$, até obter a hierarquização completa dos servidores.

2.3.2 Prova do Teorema 2.2.1

(I) Utilizando a ideia do esquema das chegadas do tipos de trabalho descrita anteriormente e assumindo para a primeira etapa da decomposição $D = C_0$ e $V_1 = \max_{C \subseteq D} W_D(C)$.

Então, para cada etapa $k = 2, 3, \ldots$ da decomposição é defina a velocidade restrita-reduzida V_k dentro de um agrupamento restrito C_k a partir do sistema restrito-reduzido $D_k = (D \setminus \cup_{n=1}^{k-1} C_n)$, ou seja,

$$V_k = \max_{C \subseteq D_k} W_{D_k}(C).$$

No entanto, se a rede estiver sendo analisada pela melhor política estática $\widetilde{\Pi}(C)$ que equaliza todos os encaminhamentos dos trabalhos para qualquer agrupamento $C \subseteq D$ do sistema restrito-reduzido D, ou seja,

$$\widetilde{\Pi}(C) = \Big\{ \pi \in \Pi^{Estat} : \sum_{i \in \mathcal{N}_D(C)} \pi(i,j) \frac{\lambda_i(D)}{\mu_{ij}} = \sum_{i \in \mathcal{N}_D(C)} \pi(i,l) \frac{\lambda_i(D)}{\mu_{il}}; \forall\, j,\ l \in C \Big\},$$

eventualmente todos os servidores que estão no sistema restrito-reduzido D da primeira etapa da decomposição terão a mesma quantidade de carga horária média total de trabalho. Desta forma, todos os servidores vão ter a mesma velocidade.

Portanto, se todos os servidores que formam o sistema restrito-reduzido D apresentam a velocidade máxima W_1, consequentemente $V_1 = W_1$. Então, todos os servidores formarão um único agrupamento C_1, ou seja, $C_0 = C_1$. Finalizando a prova da parte (I) do Teorema.

(II) & (III) Para as redes que apresentam $C_1 \neq C_0$, nota-se que é possível decompor a rede em mais do que um sistema restrito.

Para isso, defina para cada uma das etapas $k = 2, 3, \ldots$ da decomposição, a velocidade do sistema restrito V_k encontrada através da decomposição do sistema restrito-reduzido $D_k = \left(C_0 \setminus \cup_1^{k-1} C_i\right)$ dada por

$$V_k = \max_{C \subseteq D_k} W_{D_k}(C)\ , \quad \overline{C}_k = \cup \{ C \subseteq D_k : V_k = \max_{C \subseteq D_k} W_{D_k}(C) \}$$

onde $\overline{C}_k$ são os servidores que formam o sistema restrito cuja velocidade é V_k estabelecida pela k-ésima etapa.

Então, de modo análogo ao item anterior, para cada etapa da decomposição hierárquica $k = 2, 3, \ldots$ obtêm-se a velocidade do sistema restrito V_k e os servidores $C_k = \overline{C}_k$ que eventualmente formarão o agrupamento da etapa k da decomposição, até obter $C_k = D_k$, finalizando a decomposição hierárquica.

É importante notar que todos os servidores que formam um sistema restrito C_k possuem a mesma velocidade V_k e que este sistema restrito é constituído pelos servidores que apresentam a mesma velocidade $V_k = \max_{C \subseteq D_k} W_{D_k}(C)$ durante a etapa k que, por sua vez, são removidos do sistema restrito-reduzido $D_k = \left(C_0 \setminus \cup_1^{k-1} C_i\right)$ para uma próxima hierarquização, se possível.

□

2.3.3 Preliminares para a prova do Teorema 2.2.2

Para a análise da transiência ou da recorrência do processo X utilizam-se as funções de Lyapunov, cujas conclusões são estabelecidas através da Definição 9 do Capítulo 1 e dos Teoremas 2.3.1 & 2.3.2 apresentados a seguir.

Teorema 2.3.1. *Seja uma cadeia de Markov com espaço de estados* $\mathcal{A} = \{\eta_i, i \geq 0\}$*, irredutível e aperiódica, cuja posição da cadeia no tempo* t *é* ξ_t*. Então,*

(I) a cadeia de Markov é transiente se, e somente se, existe uma função positiva $f(\eta)$*,* $\eta \in \mathcal{A}$*, um número* $\epsilon > 0$ *e um conjunto finito* $A \in \mathcal{A}$ *tal que*

$$\mathbf{E}(f(\xi_{t+1}) - f(\xi_t) \mid \xi_t = \eta_m) < 0\,,\ \eta_m \notin A\,,$$
$$f(\eta_k) < \inf_{\eta_m \in A} f(\eta_m)\,,\ \textit{para pelo menos um } \eta_k \notin A\,.$$

Demonstração do Teorema 2.3.1. O item (I) deste teorema pode ser verificado através do Teorema 2.2.2 em Fayolle, Malyshev e Menshikov (1995, Capítulo 2).
□

Teorema 2.3.2. *Seja* $\{X_t\}$ *uma cadeia de Markov irredutível definida num espaço de estados contáveis* $\mathcal{S}$, $f : \mathcal{S} \to \mathbb{R}^+$ *e* $\Delta f_t = f(X_{t+1}) - f(X_t)$.

A Se existem constantes $c > 0$, $d > 0$ *e* $\epsilon > 0$ *tais que*

A1 $|\Delta f_t| < d$ *q.c. e*

A2 $\mathbf{E}(\Delta f_t \mid X_t = x) > \epsilon$ *para todo* $x \in \{x : f(x) > c\}$,

então $\{X_t\}$ *é transiente.*

B Se existe uma constante $\epsilon > 0$ *e um conjunto finito* $A \subset \mathcal{S}$ *tais que*

B1 $\mathbf{E}(f(X_{t+1}) \mid X_t = x) < \infty$ *para* $x \in A$ *e*

B2 $\mathbf{E}(\Delta f_t \mid X_t = x) \leq -\epsilon$ *para* $x \in \mathcal{S} \setminus A$,

então $\{X_t\}$ *é recorrente positivo.*

Demonstração: O item (A) deste teorema pode ser verificado através do Teorema 2.2.7 em Fayolle, Malyshev e Menshikov (1995, Capítulo 2). Enquanto que o item (B) é o Critério de Foster, podendo ser verificado pelo Teorema 2.2.4 em Fayolle, Malyshev e Menshikov (1995, Capítulo 2) ou pela Proposição I.5.3 em Asmussen (2003). □

2.3.4 Prova do Teorema 2.2.2

A demonstração do Teorema 2.2.2 consiste em provar que se $V_1 > 0$, então o processo X com respeito à carga horária média total de trabalho na fila dos servidores é transiente para qualquer política estática. Para isso, será verificada a condição (A) do Teorema 2.3.2.

Lembre-se de que as chegadas dos tipos de trabalho nos servidores $C_0 = \{1, \ldots, N\}$ são processos de Poisson independentes para qualquer política de escolha $\pi \in \Pi'$. Desta forma, a taxa de chegada dos trabalhos para o servidor j é dada por $\sum_{i \in \mathcal{N}(C_0)} \lambda_i \pi(i, j)$, para todo $j = 1, \ldots, N$.

Esta demonstração é realizada usando a cadeia X_t com os saltos $\mathbf{x}$ descritos na Seção 2.1 e a função de Lyapunov

$$L(X_t) = \sum_{j \in C_1} \left(\sum_{\zeta \in \mathcal{N}(C_1)} \mu_{\zeta j} \right) x_j \,. \tag{2.13}$$

Agora, considere a notação $\mathbf{0}$ para todo estado $x_j = 0$, referindo-se à fila do servidor j que não tem nenhum trabalho; e $\mathbf{1}$ para todo estado $x_j = 1$, referindo-se que existe pelo menos um trabalho na fila do servidor j. Então,

$$\alpha(\mathbf{0}) = \sum_{i \in \mathcal{N}(C_0)} \lambda_i \leq \alpha(\mathbf{x}) \leq \sum_{i \in \mathcal{N}(C_0)} \lambda_i + \sum_{j \in C_0} 1 = \alpha(\mathbf{1}) \ . \tag{2.14}$$

Perceba que a taxa dos eventos do processo X na configuração atual $\mathbf{x}$ dada por

$$\alpha(\mathbf{x}) = \sum_{i \in \mathcal{N}(C_0)} \lambda_i + \sum_{j \in C_0} \mathbf{1}_{\{x_j > 0\}} \text{ é limitada.}$$

Logo, para qualquer política estática π num agrupamento C_1 de servidores com $D = C_0$ e considerando a função de Lyapunov (2.13), pode-se encontrar a carga horária média total de trabalho esperada para o processo X como apresentada a seguir.

Por simplicidade da demonstração apresentada abaixo, considere a notação: $\Delta L_t = L(X_{t+1}) - L(X_t)$, $A_i = \mathbf{1}_{\{\text{ chegada em } S_i\}}$, $R_{ij} = \mathbf{1}_{\{\text{ chegada em } S_i \text{ encaminhada ao servidor } j\}}$, $D_j = \mathbf{1}_{\{\text{saída do servidor } j\}}$ e $\mathbf{E}_\pi$ o valor esperado para qualquer política estática.

$$\begin{aligned}
\mathbf{E}_\pi(\Delta L_t \mid X_t = \mathbf{x}) &= \sum_{i \in \mathcal{N}(C_1)} \sum_{j \in C_1} \mathbb{P}(A_i R_{ij}) \mathbf{E}_\pi(\Delta L_t \mid A_i R_{ij}) + \sum_{j \in C_1} \mathbb{P}(D_j) \mathbf{E}_\pi(\Delta L_t \mid D_j) \\
&= \sum_{i \in \mathcal{N}(C_1)} \sum_{j \in C_1} \lambda_i \frac{\pi(\mathbf{x}, i, j)}{\sum\limits_{i \in \mathcal{N}(C_1)} \lambda_i + N} \, \mathbf{E}_\pi(\Delta L_t \mid A_i R_{ij}) \\
&\quad + \sum_{j \in C_1} \frac{1}{\sum\limits_{i \in \mathcal{N}(C_1)} \lambda_i + N} \, \mathbf{E}_\pi(\Delta L_t \mid D_j)
\end{aligned}$$

Considere $\alpha(\mathbf{x}) = \sum\limits_{i \in \mathcal{N}(C_1)} \lambda_i + N$ e $a_j = \sum\limits_{\zeta \in \mathcal{N}(C_1)} \mu_{\zeta j}$. Então, tem-se que

$$\begin{aligned}
\alpha(\mathbf{x}) \mathbf{E}_\pi(\Delta L_t \mid X_t = \mathbf{x}) &= \sum_{i \in \mathcal{N}(C_1)} \sum_{j \in C_1} \lambda_i \pi(\mathbf{x}, i, j) \mathbf{E}_\pi(\Delta L_t \mid A_i R_{ij}) + \sum_{j \in C_1} \mathbf{E}_\pi(\Delta L_t \mid D_j) \\
&= \sum_{i \in \mathcal{N}(C_1)} \sum_{j \in C_1} \lambda_i \pi(\mathbf{x}, i, j) \Big\{ a_j \left(x_j + \frac{1}{\mu_{ij}} \right) + \sum_{l \neq j} a_l x_l \Big\} \\
&\quad + \sum_{j \in C_1} \Big\{ a_j (x_j - 1) + \sum_{l \neq j} a_l x_l \Big\} - \sum_{u \in C_1} a_u x_u
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in\mathcal{N}(C_1)} \sum_{j\in C_1} \lambda_i \pi(x,i,j)\Big\{\frac{a_j}{\mu_{ij}}\Big\} - \sum_{j\in C_1} a_j \\
&= \sum_{j\in C_1} a_j \left(\sum_{i\in\mathcal{N}(C_1)} \pi(x,i,j)\frac{\lambda_i}{\mu_{ij}} - 1 \right) \\
&= \sum_{j\in C_1} a_j V(j,x,\pi)
\end{aligned}
$$

Portanto, para todo estado $\mathbf{x} \in \mathbb{R}^N$,

$$
\alpha(x)\mathbf{E}_\pi(\Delta L_t \mid X_t = x) = \sum_{j\in C_1} a_j V(j,x,\pi) = \sum_{j\in C_1} \sum_{\zeta\in\mathcal{N}(C_1)} \mu_{\zeta j} V_1 = \sum_{\zeta\in\mathcal{N}(C_1)} \mu_{\zeta 1}|C_1|V_1 \,. \tag{2.15}
$$

Com este resultado, através do item (A) do Teorema 2.3.2, pode-se concluir que a cadeia de salto X_t correspondente ao processo X é transiente para qualquer política estática.

No entanto, se para todo estado $\mathbf{x} \in \mathbb{R}^N$ tem-se

$$
\alpha(x)\mathbf{E}_\pi(\Delta L_t \mid X_t = x) = \sum_{j\in C} a_j V(j,x,\pi) < 0. \tag{2.16}
$$

Então, através do item (B) do Teorema 2.3.2, é possível concluir que se a velocidade do primeiro agrupamento é $V_1 < 0$, conseqüentemente o processo X é recorrente positivo para a melhor política estática.

Observação 3. *Note que é possível ter a igualdade de* (2.15) *no caso das políticas estáticas que possuem a capacidade de equalizar as chegadas dos tipos de trabalho da rede. Caso contrário, se as políticas de escolha de servidor não equalizam as chegadas dos tipos de trabalho, isso significa que podem existir servidores sobrecarregados em comparação com os demais servidores. Portanto, a velocidade poderá ser igual ou maior em alguns locais da rede, dependendo da política de escolha que a rede é submetida.*

□

Refazendo o cálculo para o processo Y, obtêm-se que

$$\begin{aligned}
\alpha(\mathbf{y})\mathbf{E}_\pi(\Delta L_t \mid Y_t = \mathbf{y}) &= \sum_{i\in\mathcal{N}(C_1)}\sum_{j\in C_1}\lambda_i\pi(\mathbf{y},i,j)\mathbf{E}_\pi(\Delta L_t \mid A_i R_{ij}) + \sum_{j\in C_1}\mu_{y_1 j}\mathbf{E}_\pi(\Delta L_t \mid D_j)\\
&= \sum_{i\in\mathcal{N}(C_1)}\sum_{j\in C_1}\lambda_i\pi(\mathbf{y},i,j)\Big\{a_j\,(y_j+1)+\sum_{l\neq j}a_l y_l\Big\}\\
&\quad + \sum_{j\in C_1}\mu_{y_1 j}\Big\{a_j\,(y_j-1)+\sum_{l\neq j}a_l y_l\Big\} - \sum_{u\in C_1}a_u x_u\\
&= \sum_{i\in\mathcal{N}(C_1)}\sum_{j\in C_1}\lambda_i\pi(\mathbf{y},i,j)a_j - \sum_{j\in C_1}\mu_{y_1 j}a_j\\
&= \sum_{j\in C_1}a_j\left(\sum_{i\in\mathcal{N}(C_1)}\pi(\mathbf{y},i,j)\lambda_i - \mu_{y_1 j}\right)\\
&= \sum_{j\in C_1}a_j\mu_{y_1 j}\left(\sum_{i\in\mathcal{N}(C_1)}\frac{\pi(\mathbf{y},i,j)\lambda_i}{\mu_{y_1 j}} - 1\right),
\end{aligned}$$

onde $\mu_{y_1 j}$ é a taxa do atendimento do trabalho $y_1 \in \{1,\ldots,K\}$ que está na primeira posição da fila do servidor j.

Note que o cálculo do valor esperado para o processo Y exige uma análise muito cuidadosa e precisa das taxas dos atendimentos dos tipos de trabalho que estão na fila dos servidores e na ordem que chegaram. Enquanto que para o processo X este conhecimento das taxas dos atendimentos dos tipos de trabalho que estão na fila dos servidores e na ordem que chegaram não é necessário para o desenvolvimento dos cálculos.

A partir dos resultados simulados para dois exemplos de redes de filas apresentados no Capítulo 3 é possível perceber que o processo X apresenta um comportamento semelhante ao comportamento obtido pelo processo Y. Portanto, pode-se concluir que o processo X é uma boa opção para analisar o processo Y.

2.3.5 Preliminares para a prova do Teorema 2.2.3

O resultado do Teorema 2.2.3 com respeito ao passeio aleatório $\xi(t)$ é estabelecido através do processo X sem a exigência do processo ser positivo.

O primeiro resultado demonstrado nesta subseção apresenta a garantia de que todos os servidores que estão dentro de um mesmo agrupamento C possuem a mesma velocidade.

Para isso, considere a taxa geral para todo o estado $\mathbf{x}$ dada por

$$\alpha(\mathbf{x}) = \sum_{i \in \mathcal{N}(C_0)} \lambda_i + \sum_{j \in C_0} 1 = \sum_{i \in \mathcal{N}(C_0)} \lambda_i + N\,.$$

Lembre-se de que a união de todas as chegadas dos trabalhos de S^D oferece o fluxo total das chegadas dos trabalhos $\lambda_i(D)$ do sistema restrito-reduzido $D = \cup_{l=k}^{K} C_l$ submetido a qualquer política estática $\pi \in \Pi_K$.

Definição 38. *Um estado* $\mathbf{x}$ *é* agrupado adequadamente *quando*

$$\max_{i \in \mathcal{N}(C_n)} \left\{ \left[x_j + (1/\mu_{ij}) \right] / \mu_{ij} \right\} < \min_{i \in \mathcal{N}(C_n)} \left\{ \left[x_l + (1/\mu_{il}) \right] / \mu_{il} \right\} \text{ para cada } l \in \cup_{n=1}^{k-1} C_n \text{ e}$$

$$\max_{i \in \mathcal{N}(C_n)} \left\{ \left[x_j + (1/\mu_{ij}) \right] / \mu_{ij} \right\} > \min_{i \in \mathcal{N}(C_n)} \left\{ \left[x_l + (1/\mu_{il}) \right] / \mu_{il} \right\} \text{ para cada } l \in \cup_{n=k+1}^{K} C_n\,.$$

□

Para a demonstração do Teorema 2.2.3, de forma análoga ao teorema anterior, utilizam-se as funções de Lyapunov $F_k(\mathbf{x})$ adequadas para analisar os encaminhamentos dos trabalhos aos servidores segundo as políticas de escolha previamente estabelecidas.

Portanto, considerando o passeio aleatório $\xi(t)$ em cada agrupamento C_k de servidores e $a_j = \sum_{\zeta \in \mathcal{N}(C_k)} \mu_{\zeta j}$ para todo $j \in C_k$, as funções de Lyapunov para esta demonstração são:

(a) para a política JSWA, referente a menor carga horária média total de trabalho da fila, ponderada pelo tempo médio de processamento dos servidores que podem realizar o trabalho,

$$F_k^1(\mathbf{x}) = \sum_{r,l \in C_k} \frac{(a_r x_r - a_l x_l)^2}{2 a_r^2 a_l^2} \text{ e} \tag{2.17}$$

(b) para a política JSWAS, referente a menor carga horária média total de trabalho do sistema, ponderada pelo tempo médio de processamento dos servidores que podem realizar o trabalho,

$$F_k^2(\mathbf{x}) = \sum_{æ \in \mathcal{N}_D(C_k)} \sum_{r,l \in C_k} \frac{\left(a_r x_{(r,æ)} - a_l x_{(l,æ)}\right)^2}{2a_r^2 a_l^2}, \tag{2.18}$$

onde $x_{(j,æ)} = x_j + (1/\mu_{æj})$ para todo $j = r, l$.

Por simplicidade, considere $\Delta F_k^z(\mathbf{x}) = F_k^z(\mathbf{x}_{\eta+1}) - F_k^z(\mathbf{x}_\eta)$ para z=1,2; $\mathbf{E}_Q$ a carga horária média total de trabalho esperada para a política JSWA; $\mathbf{E}_S$ a carga horária média total de trabalho esperada para a política JSWAS e $\mathbf{E}_\pi$ a carga horária média total de trabalho esperada para a melhor política estática.

Lema 2.3.1. *Considerando uma cadeia imersa* ξ_η, $\gamma_k = \sum_{l \in C_k} \left(1/a_l^2\right)$ *e os processos* $F_k^1(\mathbf{x})$ *e* $F_k^2(\mathbf{x})$. *Então,*

(a) para qualquer política estática $\pi \in \Pi_k$,

(a1) segundo a função de Lyapunov para a política JSWA, tem-se

$$\alpha(\mathbf{x})\mathbf{E}_\pi\left(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right) = \gamma_k \sum_{j \in C_k} \left(\sum_{i \in \mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2} + 1 \right);$$

(a2) segundo a função de Lyapunov para a política JSWAS, tem-se

$$\alpha(\mathbf{x})\mathbf{E}_\pi\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right) = \gamma_k |\mathcal{N}_D(C_k)| \sum_{j \in C_k} \left(\sum_{i \in \mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2} + 1 \right);$$

(b) para qualquer estado $\mathbf{x}$ *agrupado adequadamente para a política JSWA, tem-se*

$$\begin{aligned} \alpha(\mathbf{x})\mathbf{E}_Q\left(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta\right) &= 2\gamma_k \sum_{i \in \mathcal{N}_D(C_k)} \lambda_i(D)\left[\underline{x_i} - \sum_{j \in C_k} \pi(i,j)\frac{x_j}{\mu_{ij}}\right] \\ &\quad + \gamma_k \sum_{j \in C_k} \left(\sum_{i \in \mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(\mathbf{x},i,j)}{\mu_{ij}^2} + 1 \right), \end{aligned}$$

quando $\underline{x_i} \leq \sum_{j \in C_k} \pi(i,j)\left(x_j/\mu_{ij}\right)$ *para cada* $i \in \mathcal{N}_D(C_k)$ *em qualquer estado* $\mathbf{x}$.

(c) para qualquer estado $\mathbf{x}$ *agrupado adequadamente para a política JSWAS, tem-se*

$$\alpha(\mathbf{x})\mathbf{E}_S\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta\right) = 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D) \sum_{æ\in\mathcal{N}_D(C_k)} \left[\underline{x}_i^* - \sum_{j\in C_k} \pi(i,j)\frac{x_{(j,æ)}}{\mu_{ij}}\right] + \gamma_k|\mathcal{N}_D(C_k)| \sum_{j\in C_k}\left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(\mathbf{x},i,j)}{\mu_{ij}^2} + 1\right),$$

quando $\underline{x}_i^* \leq \sum_{j\in C_k} \pi(i,j)\left(x_{(j,æ)}/\mu_{ij}\right)$ *para cada* $i \in \mathcal{N}_D(C_k)$ *em qualquer estado* $\mathbf{x}$.

Observação 4. *As políticas estáticas não consideram a configuração atual da rede no momento dos encaminhamentos dos trabalhos aos servidores. No entanto, a carga horária média total de trabalho esperada para as Lyapunov* $F_k^1(\mathbf{x})$ *e* $F_k^2(\mathbf{x})$ *são encontradas para estas políticas porque estes resultados são úteis para obter os resultados para as políticas dinâmicas, ou seja, para as políticas JSWA e JSWAS, contribuindo para as comparações das políticas analisadas em questão.*

Demonstração do Lema 2.3.1: (a) Para qualquer política estática $\pi \in \Pi_K$, a velocidade das filas que estão dentro do agrupamento C_k é dada por

$$V_k = \sum_{i\in\mathcal{N}_D(C_k)} \pi(i,j)\frac{\lambda_i(D)}{\mu_{ij}} - 1\ .$$

Portanto, se um trabalho chegou à rede e este foi encaminhado para a fila do servidor j, então existirá um aumento em média de $(1/\mu_{ij})$ na carga horária média total de trabalho da fila deste servidor j, ou seja, $\mathbf{x} = (x_1, \ldots, x_j + 1/\mu_{ij}, \ldots, x_n)$.

No entanto, se o servidor j processou completamente um trabalho de carga horária média de $(1/\mu_{ij})$, isto significa que houve uma saída, em média, de um trabalho da fila do servidor j, cuja taxa de atendimento é de μ_{ij}, ou seja, $\mathbf{x} = (x_1, \ldots, x_j - 1, \ldots, x_n)$.

Com base nisso, é possível obter o resultado do valor esperado do item (a1) para a política JSWA como apresentado a seguir.

$$
\begin{aligned}
&\mathbf{E}_\pi(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}) \\
&= 2\Big\{ \sum_{i\in\mathcal{N}_D(C_k)} \sum_{j\in C_k} \frac{\lambda_i(D)\pi(i,j)}{\alpha(\mathbf{x})} \Big[\sum_{l\neq j\in C_k} \frac{1}{2a_l^2 a_j^2} \left(2a_j x_j \frac{a_j}{\mu_{ij}} + \frac{a_j^2}{\mu_{ij}^2} \right) \Big] \\
&\quad + \sum_{j\in C_k} \frac{1}{\alpha(\mathbf{x})} \Big[\sum_{l\neq j\in C_k} \frac{1}{2a_l^2 a_j^2} \left(-2a_j^2 x_j + a_j^2 \right) \Big] \Big\} \\
&= 2\Big\{ \sum_{i\in\mathcal{N}_D(C_k)} \sum_{j\in C_k} \frac{\lambda_i(D)\pi(i,j)}{\alpha(\mathbf{x})} \Big[\sum_{l\neq j\in C_k} \frac{1}{2a_l^2 a_j^2} \left(2a_j^2 x_j \frac{1}{\mu_{ij}} + \frac{a_j^2}{\mu_{ij}^2} \right) \Big] \\
&\quad + \sum_{j\in C_k} \frac{1}{\alpha(\mathbf{x})} \Big[\sum_{l\neq j\in C_k} \frac{1}{2a_l^2 a_j^2} \left(-2a_j^2 x_j + a_j^2 \right) \Big] \Big\} \\
&= 2\Big\{ \sum_{i\in\mathcal{N}_D(C_k)} \sum_{j\in C_k} \frac{\lambda_i(D)\pi(i,j)}{\alpha(\mathbf{x})} \sum_{l\in C_k} \frac{1}{a_l^2} \Big[\frac{x_j}{\mu_{ij}} + \frac{1}{2\mu_{ij}^2} \Big] \\
&\quad + \sum_{j\in C_k} \frac{1}{\alpha(\mathbf{x})} \sum_{l\in C_k} \frac{1}{a_l^2} \Big[-x_j + \frac{1}{2} \Big] \Big\} \\
&= \frac{2}{\alpha(\mathbf{x})} \Big\{ \sum_{j\in C_k} \gamma_k x_j \left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}} - 1 \right) + \frac{1}{2} \sum_{j\in C_k} \gamma_k \left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2} + 1 \right) \Big\} \\
&= \frac{1}{\alpha(\mathbf{x})} \Big\{ 2 \sum_{j\in C_k} \gamma_k x_j \left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}} - 1 \right) + \sum_{j\in C_k} \gamma_k \left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2} + 1 \right) \Big\}
\end{aligned}
$$

Logo, pode-se escrever

$$
\begin{aligned}
&\alpha(\mathbf{x})\mathbf{E}_\pi(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}) \\
&= 2 \sum_{j\in C_k} \gamma_k x_j \left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}} - 1 \right) + \sum_{j\in C_k} \gamma_k \left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2} + 1 \right) \\
&= \sum_{j\in C_k} 2\gamma_k x_j V_k + \gamma_k \sum_{j\in C_k} \left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2} + 1 \right) \\
&= 2\gamma_k V_k \sum_{j\in C_k} x_j + \gamma_k \sum_{j\in C_k} \left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2} + 1 \right)
\end{aligned}
$$

Note que para o vetor $\widetilde{\mathbf{x}} = (1/a_1, 1/a_2, \ldots, 1/a_{|C_k|})$ e para qualquer $c \in \mathbb{R}$, se $F_k^1(\mathbf{x}) = F_k^1(\mathbf{x} + c\widetilde{\mathbf{x}})$, então é possível transladar o vetor $\mathbf{x}$ em direção ao vetor $\widehat{\mathbf{x}} = \mathbf{x} + c\widetilde{\mathbf{x}}$ de tal forma que $\sum_{j \in C_k} x_j = 0$.

Portanto, assumindo esta informação da translação do vetor $\mathbf{x}$ obtêm-se que

$$\alpha(\mathbf{x})\mathbf{E}_\pi(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}) = \gamma_k \sum_{j \in C_k} \left(\sum_{i \in \mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2} + 1 \right).$$

De forma análoga, o resultado do item $(a2)$ da política JSWAS é encontrado, como apresentado abaixo.

$$\begin{aligned}
&\mathbf{E}_\pi(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta = \mathbf{x}) \\
&= 2\Big\{ \sum_{i \in \mathcal{N}_D(C_k)} \sum_{j \in C_k} \frac{\lambda_i(D)\pi(i,j)}{\alpha(\mathbf{x})} \Big[\sum_{l \neq j \in C_k} \frac{1}{2a_l^2 a_j^2} \sum_{æ \in \mathcal{N}_D(C_k)} \left(2a_j x_{(j,æ)} \frac{a_j}{\mu_{ij}} + \frac{a_j^2}{\mu_{ij}^2} \right) \Big] \\
&\quad + \sum_{j \in C_k} \frac{1}{\alpha(\mathbf{x})} \Big[\sum_{l \neq j \in C_k} \frac{1}{2a_l^2 a_j^2} \sum_{æ \in \mathcal{N}_D(C_k)} \left(-2a_j^2 x_{(j,æ)} + a_j^2 \right) \Big] \Big\} \\
&= 2\Big\{ \sum_{i \in \mathcal{N}_D(C_k)} \sum_{j \in C_k} \frac{\lambda_i(D)\pi(i,j)}{\alpha(\mathbf{x})} \Big[\sum_{l \neq j \in C_k} \frac{1}{2a_l^2 a_j^2} \sum_{æ \in \mathcal{N}_D(C_k)} \left(2a_j^2 x_{(j,æ)} \frac{1}{\mu_{ij}} + \frac{a_j^2}{\mu_{ij}^2} \right) \Big] \\
&\quad + \sum_{j \in C_k} \frac{1}{\alpha(\mathbf{x})} \Big[\sum_{l \neq j \in C_k} \frac{1}{2a_l^2 a_j^2} \sum_{æ \in \mathcal{N}_D(C_k)} \left(-2a_j^2 x_{(j,æ)} + a_j^2 \right) \Big] \Big\} \\
&= 2\Big\{ \sum_{i \in \mathcal{N}_D(C_k)} \sum_{j \in C_k} \frac{\lambda_i(D)\pi(i,j)}{\alpha(\mathbf{x})} \gamma_k \sum_{æ \in \mathcal{N}_D(C_k)} \Big[\frac{x_{(j,æ)}}{\mu_{ij}} + \frac{1}{2\mu_{ij}^2} \Big] \\
&\quad + \sum_{j \in C_k} \frac{1}{\alpha(\mathbf{x})} \gamma_k \sum_{æ \in \mathcal{N}_D(C_k)} \Big[-x_{(j,æ)} + \frac{1}{2} \Big] \Big\} \\
&= \frac{1}{\alpha(\mathbf{x})} \Big\{ \sum_{i \in \mathcal{N}_D(C_k)} \sum_{j \in C_k} \lambda_i(D)\pi(i,j)\gamma_k \sum_{æ \in \mathcal{N}_D(C_k)} \Big[\frac{2x_{(j,æ)}}{\mu_{ij}} + \frac{1}{\mu_{ij}^2} \Big] \\
&\quad + \sum_{j \in C_k} \gamma_k \sum_{æ \in \mathcal{N}_D(C_k)} \Big[-2x_{(j,æ)} + 1 \Big] \Big\} \\
&= \frac{1}{\alpha(\mathbf{x})} \Big\{ 2 \sum_{j \in C_k} \gamma_k \sum_{æ \in \mathcal{N}_D(C_k)} x_{(j,æ)} \left(\sum_{i \in \mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}} - 1 \right)
\end{aligned}$$

$$+\sum_{j\in C_k}\gamma_k\sum_{æ\in\mathcal{N}_D(C_k)}\left(\sum_{i\in\mathcal{N}_D(C_k)}\frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2}+1\right)\Big\}$$

$$=\frac{1}{\alpha(\mathbf{x})}\Big\{2\gamma_k\sum_{j\in C_k}\sum_{æ\in\mathcal{N}_D(C_k)}x_{(j,æ)}V_k$$

$$+\sum_{j\in C_k}\gamma_k\sum_{æ\in\mathcal{N}_D(C_k)}\left(\sum_{i\in\mathcal{N}_D(C_k)}\frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2}+1\right)\Big\}$$

Usando novamente a translação do vetor $\mathbf{x}$ para o vetor $\widehat{\mathbf{x}}=\mathbf{x}+c\widetilde{\mathbf{x}}$ de tal forma que $\sum_{j\in C_k}x_{(j,æ)}=0$, para o vetor $\widetilde{\mathbf{x}}=(1/a_1,1/a_2,\ldots,1/a_{|C_k|})$ e qualquer $c\in\mathbb{R}$, obtêm-se que

$$\alpha(\mathbf{x})\mathbf{E}_\pi(\Delta F_k^2(\mathbf{x})\mid\xi_\eta=\mathbf{x})$$

$$=2\gamma_kV_k\sum_{j\in C_k}\sum_{æ\in\mathcal{N}_D(C_k)}x_{(j,æ)}+\gamma_k|\mathcal{N}_D(C_k)|\sum_{j\in C_k}\left(\sum_{i\in\mathcal{N}_D(C_k)}\frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2}+1\right)$$

$$=\gamma_k|\mathcal{N}_D(C_k)|\sum_{j\in C_k}\left(\sum_{i\in\mathcal{N}_D(C_k)}\frac{\lambda_i(D)\pi(i,j)}{\mu_{ij}^2}+1\right).$$

(b) Agora, assumindo a política JSWA é calculado $\mathbf{E}_Q\Big(\Delta F_k^1(\mathbf{x})\mid\xi_\eta=\mathbf{x}\Big)$.

Primeiramente, considere que as filas são agrupadas adequadamente para qualquer política dinâmica, de tal forma que os trabalhos $\mathcal{N}_D(C_k)$ que chegam à rede são encaminhados para os servidores que estão dentro do agrupamento C_k do sistema restrito-reduzido D.

Para evitar as repetições nos cálculos para valor esperado da política JSWA, os resultados encontrados no item (a1) são utilizados durante o desenvolvimento dos cálculos de $\mathbf{E}_Q\Big(\Delta F_k^1(\mathbf{x})\mid\xi_\eta=\mathbf{x}\Big)$.

Portanto, com base no resultado do item (a1), o valor esperado para a política JSWA é encontrado a partir da igualdade

$$\begin{aligned}\alpha(\mathbf{x})\mathbf{E}_Q\big(\Delta F_k^1(\mathbf{x})\mid\xi_\eta\big) &= \alpha(\mathbf{x})\mathbf{E}_\pi\big(\Delta F_k^1(\mathbf{x})\mid\xi_\eta\big)\\ &\quad+\alpha(\mathbf{x})\left(\mathbf{E}_Q\big(\Delta F_k^1(\mathbf{x})\mid\xi_\eta\big)-\mathbf{E}_\pi\big(\Delta F_k^1(\mathbf{x})\mid\xi_\eta\big)\right).\end{aligned}$$

Desta forma, analisando a igualdade acima, observa-se que

$$\begin{aligned}
&\alpha(\mathbf{x})\Big[\mathbf{E}_{Q}\big(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\big) - \mathbf{E}_{\pi}\big(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\big)\Big] \\
&= 2\gamma_k \sum_{j\in C_k} \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\Big[\pi(\mathbf{x},i,j) - \pi(i,j)\Big]\frac{x_j}{\mu_{ij}} \qquad (2.19)\\
&\quad + \sum_{j\in C_k} \gamma_k\Big[\sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\left(\pi(\mathbf{x},i,j) - \pi(i,j)\right)\frac{1}{\mu_{ij}^2} + 1\Big] \qquad (2.20)
\end{aligned}$$

Note que a segunda parte (2.20) do termo acima,

$$\begin{aligned}
&\sum_{j\in C_k} \gamma_k\Big[\sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\left(\pi(\mathbf{x},i,j) - \pi(i,j)\right)\frac{1}{\mu_{ij}^2} + 1\Big] \\
&= \sum_{j\in C_k} \gamma_k\Big[\sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\left(\frac{\mathbf{1}_{\{j\in\mathcal{Q}_i(\mathbf{x})\}}}{|\mathcal{Q}_i(\mathbf{x})|} - \pi(i,j)\right)\frac{1}{\mu_{ij}^2} + 1\Big] \le \pounds\,,
\end{aligned}$$

é limitada e não depende de $\mathbf{x}$.

Agora, analisando a primeira parte (2.19) e assumindo $\underline{x_i} = \min_{j\in C_k}(x_j/\mu_{ij})$ e $\sum_{j\in C_k} \mathbf{1}_{\{j\in\mathcal{Q}_i(\mathbf{x})\}} = |\mathcal{Q}_i(\mathbf{x})|$, é possível demonstrar que esta parte é negativa como se apresenta a seguir.

$$\begin{aligned}
&2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \sum_{j\in C_k} \lambda_i(D)\Big[\pi(\mathbf{x},i,j) - \pi(i,j)\Big]\frac{x_j}{\mu_{ij}} \\
&\quad = \quad 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \sum_{\substack{j\in C_k \text{ tal que}\\ x_j = \min_{l\in C_k}\left(\frac{x_l}{\mu_{il}}\right)}} \lambda_i(D)\Big[\frac{1}{|\mathcal{Q}_i(\mathbf{x})|} - \pi(i,j)\Big]\frac{x_j}{\mu_{ij}} \\
&\qquad\qquad + 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \sum_{\substack{j\in C_k \text{ tal que}\\ x_j \neq \min_{l\in C_k}\left(\frac{x_l}{\mu_{il}}\right)}} \lambda_i(D)\Big[0 - \pi(i,j)\Big]\frac{x_j}{\mu_{ij}} \\
&\quad = \quad 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\Big[\sum_{j\in C_k} \frac{\mathbf{1}_{\{j\in\mathcal{Q}_i(\mathbf{x})\}}}{|\mathcal{Q}_i(\mathbf{x})|} - \sum_{j\in C_k} \pi(i,j)\Big]\frac{x_j}{\mu_{ij}} \\
&\quad = \quad 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\Big[\underline{x_i} - \sum_{j\in C_k} \pi(i,j)\frac{x_j}{\mu_{ij}}\Big] \le 0
\end{aligned}$$

Então, combinando este resultado com a parte $(a1)$ dada pela política estática, ou seja, $\alpha(\mathbf{x})\mathbf{E}_\pi\left(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right)$, pode-se concluir que

$$\begin{aligned}\alpha(\mathbf{x})\mathbf{E}_Q\left(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta\right) &= \alpha(\mathbf{x})\mathbf{E}_\pi\left(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta\right) \\ &\quad + \alpha(\mathbf{x})\left(\mathbf{E}_Q\left(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta\right) - \mathbf{E}_\pi\left(\Delta F_k^1(\mathbf{x}) \mid \xi_\eta\right)\right) \\ &= 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\left[\underline{x_i} - \sum_{j\in C_k} \pi(i,j)\frac{x_j}{\mu_{ij}}\right] \\ &\quad + \gamma_k \sum_{j\in C_k}\left(\sum_{i\in\mathcal{N}_D(C_k)} \frac{\lambda_i(D)\pi(\mathbf{x},i,j)}{\mu_{ij}^2} + 1\right),\end{aligned}$$

quando $\underline{x_i} \leq \sum_{j\in C_k} \pi(i,j)\,(x_j/\mu_{ij})$ para cada $i \in \mathcal{N}_D(C_k)$ em qualquer estado $\mathbf{x}$.

(c) De modo análogo ao item (b) é encontrado o valor esperado para a política JSWAS, ou seja, $\mathbf{E}_S\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right)$.

Então, considerando novamente que as filas são agrupadas adequadamente para qualquer política dinâmica, de tal forma que os trabalhos $\mathcal{N}_D(C_k)$ que chegam à rede são encaminhados para os servidores que estão dentro do agrupamento C_k do sistema restrito-reduzido D.

Portanto, com base no resultado do item $(a2)$, o valor esperado para a política JSWAS é encontrado a partir da igualdade

$$\begin{aligned}\alpha(\mathbf{x})\mathbf{E}_S\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta\right) &= \alpha(\mathbf{x})\mathbf{E}_\pi\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta\right) \\ &\quad + \alpha(\mathbf{x})\left(\mathbf{E}_S\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta\right) - \mathbf{E}_\pi\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta\right)\right).\end{aligned}$$

Desta forma, analisando a igualdade acima, observa-se que

$$\begin{aligned}&\alpha(\mathbf{x})\left[\mathbf{E}_S\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right) - \mathbf{E}_\pi\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right)\right] \\ &= 2\gamma_k \sum_{j\in C_k}\sum_{æ\in\mathcal{N}_D(C_k)}\left[\sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\left(\pi(\mathbf{x},i,j) - \pi(i,j)\right)\frac{x_{(j,æ)}}{\mu_{ij}}\right] \qquad (2.21) \\ &\quad + \gamma_k \sum_{j\in C_k}\sum_{æ\in\mathcal{N}_D(C_k)}\left[\sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\left(\pi(\mathbf{x},i,j) - \pi(i,j)\right)\frac{1}{\mu_{ij}^2} + 1\right]. \qquad (2.22)\end{aligned}$$

Novamente, note que a segunda parte (2.22) do termo acima,

$$\begin{aligned}
&\gamma_k \sum_{j\in C_k} \sum_{æ\in\mathcal{N}_D(C_k)} \Big[\sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\left(\pi(\mathbf{x},i,j)-\pi(i,j)\right)\frac{1}{\mu_{ij}^2}+1\Big] \\
&= \gamma_k \sum_{j\in C_k} \sum_{æ\in\mathcal{N}_D(C_k)} \Big[\sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D)\left(\frac{\mathbf{1}_{\{j\in\mathcal{B}_i(\mathbf{x})\}}}{|\mathcal{B}_i(\mathbf{x})|}-\pi(i,j)\right)\frac{1}{\mu_{ij}^2}+1\Big] \leq ¥\,,
\end{aligned}$$

é limitada e não depende de x.

Para a análise da primeira parte (2.21) apresentada acima, considerando $\underline{x}_i^* = \min_{j\in C_k}\left(x_{(j,æ)}/\mu_{ij}\right)$, onde $x_{(j,æ)} = x_j + (1/\mu_{æj})$ para todo $j \in C_k$, $æ \in \mathcal{N}_D(C_k)$ e $\mathbf{1}_{\{j\in\mathcal{B}_i(\mathbf{x})\}} = |\mathcal{B}_i(\mathbf{x})|$, tem-se

$$\begin{aligned}
&2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D) \sum_{æ\in\mathcal{N}_D(C_k)} \sum_{j\in C_k} \Big[\pi(\mathbf{x},i,j)-\pi(i,j)\Big]\frac{x_{(j,æ)}}{\mu_{ij}} \\
&= 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D) \sum_{æ\in\mathcal{N}_D(C_k)} \sum_{\substack{j\in C_k \text{ tal que} \\ x_{(j,æ)}=\min_{l\in C_k}\left(\frac{x_{(l,æ)}}{\mu_{il}}\right)}} \Big[\frac{1}{|\mathcal{B}_i(\mathbf{x})|}-\pi(i,j)\Big]\frac{x_{(j,æ)}}{\mu_{ij}} \\
&\quad + 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D) \sum_{æ\in\mathcal{N}_D(C_k)} \sum_{\substack{j\in C_k \text{ tal que} \\ x_{(j,æ)}\neq\min_{l\in C_k}\left(\frac{x_{(l,æ)}}{\mu_{il}}\right)}} \Big[0-\pi(i,j)\Big]\frac{x_{(j,æ)}}{\mu_{ij}} \\
&= 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D) \sum_{æ\in\mathcal{N}_D(C_k)} \Big[\sum_{j\in C_k}\frac{\mathbf{1}_{\{j\in\mathcal{B}_i(\mathbf{x})\}}}{|\mathcal{B}_i(\mathbf{x})|}-\sum_{j\in C_k}\pi(i,j)\Big]\frac{x_{(j,æ)}}{\mu_{ij}} \\
&= 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D) \sum_{æ\in\mathcal{N}_D(C_k)} \Big[\underline{x}_i^* - \sum_{j\in C_k}\pi(i,j)\frac{x_{(j,æ)}}{\mu_{ij}}\Big] \leq 0\,.
\end{aligned}$$

Então, combinando este resultado com a parte (a2) dada pela política estática, ou seja, $\alpha(\mathbf{x})\mathbf{E}_\pi\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right)$, pode-se concluir que

$$\begin{aligned}
\alpha(\mathbf{x})\mathbf{E}_S\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta\right) &= \alpha(\mathbf{x})\mathbf{E}_\pi\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta\right) \\
&\quad + \alpha(\mathbf{x})\left(\mathbf{E}_S\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta\right) - \mathbf{E}_\pi\left(\Delta F_k^2(\mathbf{x}) \mid \xi_\eta\right)\right) \\
&= 2\gamma_k \sum_{i\in\mathcal{N}_D(C_k)} \lambda_i(D) \sum_{æ\in\mathcal{N}_D(C_k)} \Big[\underline{x}_i^* - \sum_{j\in C_k}\pi(i,j)\frac{x_{(j,æ)}}{\mu_{ij}}\Big]
\end{aligned}$$

$$+\gamma_k|\mathcal{N}_D(C_k)|\sum_{j\in C_k}\left(\sum_{i\in\mathcal{N}_D(C_k)}\frac{\lambda_i(D)\pi(x,i,j)}{\mu_{ij}^2}+1\right),$$

quando $\underline{x}_i^* \leq \sum_{j\in C_k}\pi(i,j)\left(x_{(j,æ)}/\mu_{ij}\right)$ para cada $i \in \mathcal{N}_D(C_k)$ em qualquer estado x. □

No entanto, é possível notar que com base nestes resultados encontrados acima, ainda é difícil de perceber qual das duas políticas dinâmicas é a política que apresenta um desempenho melhor para o modelo de redes de filas analisado.

Esta conclusão torna-se mais fácil de ser percebida através dos resultados obtidos através das simulações de dois exemplos de redes de filas apresentados no Capítulo 3. Através destes resultados simulados é possível concluir que a política JSWA apresenta um desempenho igual ao desempenho obtido pela política JSWAS. No entanto, estas políticas não apresentam um desempenho tão eficiente quanto o desempenho apresentado pela política JFQ_b.

2.3.6 Prova do Teorema 2.2.3

Se $|C| = 1$, então não existe nada a fazer. Portanto, suponha que $|C| \geq 2$ e considere que os servidores que estão dentro do agrupamento $C \subseteq C_k$ sejam conectados, ou seja, que existe uma política $\pi \in \Pi_k$ de tal forma que para quaisquer servidores j, $m \in C$, existe um encaminhamento dos trabalhos i para algum servidor j de uma rede $G(\pi)$ com vértices $\mathcal{N}_D(C) \cup C$ e elos $\{(S_i, j) : S_i \in \mathcal{N}_D(C), j \in C$ e $\pi(i,j) > 0\}$.

Como o agrupamento C é finito, então existe $\epsilon > 0$ de tal forma que a probabilidade de qualquer encaminhamento dos trabalhos de tipo i aos servidores C é dada por $\pi(i,j) \geq \epsilon$. Além disso, considere a existência de λ^- de tal maneira que as chegadas dos trabalhos do tipo i são $\lambda_i(D) \geq \lambda^-$ para todo $i : S_i \in \mathcal{N}_D(C)$.

Para realizar a demonstração deste resultado é importante considerar as funções de Lyapunov estabelecidas para cada uma das políticas dinâmicas dentro de um agrupamento C de servidores, ou seja,

(a) para a política JSWA,

$$F_C^1(\mathbf{x}) = \sum_{r,l \in C} \frac{(a_r x_r - a_l x_l)^2}{2a_r^2 a_l^2} \text{ e}$$

(b) para a política JSWAS,

$$F_C^2(\mathbf{x}) = \sum_{æ \in \mathcal{N}_D(C)} \sum_{r,l \in C_1} \frac{\left(a_r x_{(r,æ)} - a_l x_{(l,æ)}\right)^2}{2a_r^2 a_l^2},$$

onde $a_j = \sum_{\zeta \in \mathcal{N}(C)} \mu_{\zeta j}$ para todo $j \in C$ e $x_{(j,æ)} = x_j + (1/\mu_{æj})$ para todo $j = r, l$.

Desta forma, considerando a função de Lyapunov $F_C^1(\mathbf{x})$ e repetindo os cálculos do Lema 2.3.1 para um agrupamento C de servidores, obtêm-se

$$\alpha(\mathbf{x})\mathbf{E}_Q\left(\Delta F_C^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right) = 2\gamma_c \sum_{i \in \mathcal{N}_D(C)} \lambda_i(D)\left[\underline{x}_i - \sum_{j \in C} \pi(i,j)\frac{x_j}{\mu_{ij}}\right] \tag{2.23}$$

$$+ \gamma_c \sum_{j \in C}\left(\sum_{i \in \mathcal{N}_D(C)} \frac{\lambda_i(D)\pi(\mathbf{x},i,j)}{\mu_{ij}^2} + 1\right). \tag{2.24}$$

Como as taxas dos eventos são limitadas para as configurações $\mathbf{x}$ da rede, então a segunda parte da igualdade de $\alpha(\mathbf{x})\mathbf{E}_Q\left(\Delta F_C^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right)$ dada por (2.24) será

$$\gamma_c \sum_{j \in C}\left(\sum_{i \in \mathcal{N}_D(C)} \frac{\lambda_i(D)\pi(\mathbf{x},i,j)}{\mu_{ij}^2} + 1\right) \leq \mathcal{A}$$

para qualquer constante $\mathcal{A}$.

Pelo fato de que existe um $j \in C$ tal que $\underline{x}_i - x_j < -\mathcal{H}$, onde $\mathcal{H}$ pode ser escolhido arbitrariamente grande, faz com que a primeira parte da igualdade de $\alpha(\mathbf{x})\mathbf{E}_Q\left(\Delta F_C^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right)$ dada por (2.23) seja negativa, de tal maneira que se pode obter $\alpha(\mathbf{x})\mathbf{E}_Q\left(\Delta F_C^1(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right) < 0$, como no item (B1) do Teorema 2.3.2.

Agora, assumindo que $\hat{w} = \max_{j \in C} a_j$ e observando que se $F^1_C(\xi_\eta) > M^2|C|^2\hat{w}^2$, então $(a_j x_j - a_l x_l) > M$ para algum par de servidores j, $l \in C$. Suponha que $(a_j x_j - a_l x_l) > M$ para algum par de servidores $j, l \in C$. Como o agrupamento C de servidores é limitado, então existe um encaminhamento do trabalho i para um servidor j numa rede bipartida $G(\pi)$. Os encaminhamentos em $G(\pi)$ têm seus vértices alternadamente em $\mathcal{N}_D(C)$ (tipos de trabalho) e em C (servidores). Então, deve existir um encaminhamento (r, S_i, r'), tais que r, $r' \in S_i$ e $(a_{r'} x_{r'} - a_r x_r) > M/(|C| - 1)$.

Desta forma, considerando $M \leq \mathcal{A}(|C| - 1)/2\gamma_c\lambda^-\epsilon$ e toda a configuração $\mathbf{x} \in \mathbb{R}^{|C|}$ tal que para todo $i, j \in C$ tem-se $|x_i - x_j| < \mathcal{J}$, obtêm-se que

$$\mathbf{E}_Q\left(\Delta F^1_C(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right) \leq \frac{2\gamma_c\lambda^-\epsilon M}{(|C| - 1)} + \mathcal{A},$$

como no item (B2) do Teorema 2.3.2.

Portanto, através do item (II) do Teorema 2.3.2, pode-se concluir que o processo $F^1_C(\mathbf{x})$ é recorrente positivo.

De forma análoga à apresentada acima, considerando a função $F^2_C(\mathbf{x})$, obtêm-se que

$$\alpha(\mathbf{x})\mathbf{E}_S\left(\Delta F^2_C(\mathbf{x}) \mid \xi_\eta\right) = 2\gamma_C \sum_{i \in \mathcal{N}_D(C)} \lambda_i(D) \sum_{æ \in \mathcal{N}_D(C)} \left[\underline{x}^*_i - \sum_{j \in C} \pi(i, j) \frac{x_{(j,æ)}}{\mu_{ij}}\right] \quad (2.25)$$

$$+ \gamma_C|\mathcal{N}_D(C)| \sum_{j \in C} \left(\sum_{i \in \mathcal{N}_D(C)} \frac{\lambda_i(D)\pi(\mathbf{x}, i, j)}{\mu^2_{ij}} + 1 \right). \quad (2.26)$$

Como as taxas dos eventos são limitadas para as configurações $\mathbf{x}$ da rede, então a segunda parte da igualdade de $\alpha(\mathbf{x})\mathbf{E}_S\left(\Delta F^2_C(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right)$ dada por (2.26) será

$$\gamma_C|\mathcal{N}_D(C)| \sum_{j \in C} \left(\sum_{i \in \mathcal{N}_D(C)} \frac{\lambda_i(D)\pi(\mathbf{x}, i, j)}{\mu^2_{ij}} + 1 \right) \leq \mathcal{E}$$

para qualquer constante $\mathcal{E}$.

Pelo fato de que existe um $j \in C$ tal que $\underline{x}_i^* - x_j < -\mathcal{H}$, onde $\mathcal{H}$ pode ser escolhido arbitrariamente grande, faz com que a primeira parte da igualdade de $\alpha(\mathbf{x})\mathbf{E}_S\left(\Delta F_C^2(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right)$ dada por (2.25) seja negativa, de tal maneira que se pode obter $\alpha(\mathbf{x})\mathbf{E}_S\left(\Delta F_C^2(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right) < 0$, como no item (B1) do Teorema 2.3.2.

Assumindo novamente que $\hat{w} = \max_{j \in C} a_j$ e observando que se $F_C^2(\xi_\eta) > M^2|C|^2\hat{w}^2$, então $(a_j x_j - a_l x_l) > M$ para algum par de servidores $j, l \in C$. Suponha que $(a_j x_j - a_l x_l) > M$ para algum par de servidores $j, l \in C$. Como o agrupamento C de servidores é limitado, então existe um encaminhamento do trabalho i para um servidor j numa rede bipartida $G(\pi)$. Os encaminhamentos em $G(\pi)$ têm seus vértices alternadamente em $\mathcal{N}_D(C)$ (tipos de trabalho) e em C (servidores). Então, deve existir um encaminhamento tal que $r, r' \in S_i$ e $(a_{r'} x_{r'} - a_r x_r) > M/(|C|-1)$.

Portanto, considerando $M \leq \mathcal{E}(|C|-1)/2\gamma_c\lambda^-\epsilon$ e toda a configuração $\mathbf{x} \in \mathbb{R}^{|C|}$ tal que para todo $i, j \in C$ onde $|x_i - x_j| < \mathcal{J}$, obtêm-se que

$$\mathbf{E}_S\left(\Delta F_C^2(\mathbf{x}) \mid \xi_\eta = \mathbf{x}\right) \leq \frac{2\gamma_c\lambda^-\epsilon M}{(|C|-1)} + \mathcal{E},$$

como no item (B2) do Teorema 2.3.2.

Desta forma, através do item (II) do Teorema 2.3.2, pode-se concluir que o processo $F_C^2(\mathbf{x})$ também é recorrente positivo.

□

Capítulo 3

Simulações

Este capítulo descreve as ideias principais dos algoritmos dos encaminhamentos dos trabalhos aos servidores da rede para cada uma das políticas de escolha analisadas. Em seguida, os resultados e as conclusões das simulações para os processos Y, X e ξ são apresentados.

Estes resultados simulados complementam os resultados teóricos exibidos no Capítulo 2 e indicam que o processo X realmente apresenta um comportamento semelhante ao comportamento apresentado pelo processo Y.

3.1 Descrição do modelo simulado

Considere um modelo de rede que possui a capacidade de atender 3 tipos de trabalho e que apresenta 3 servidores que podem realizar os processamentos dos trabalhos que chegam à rede.

No entanto, para cada trabalho i, $i = 1, 2$ e 3, que chega à rede de filas existem somente 2 servidores que apresentam a capacidade de realizar o processamento (veja a Figura 3.1), cujas taxas dos atendimentos são definidas por

$$\mu_{ij} = \begin{cases} \mu_r & \text{se j=i (servidor j com taxa de atendimento rápida para o trabalho i)} \\ \mu_l & \text{se j}\neq i \quad \text{(servidor j com taxa de atendimento lenta para o trabalho i)} \end{cases}$$

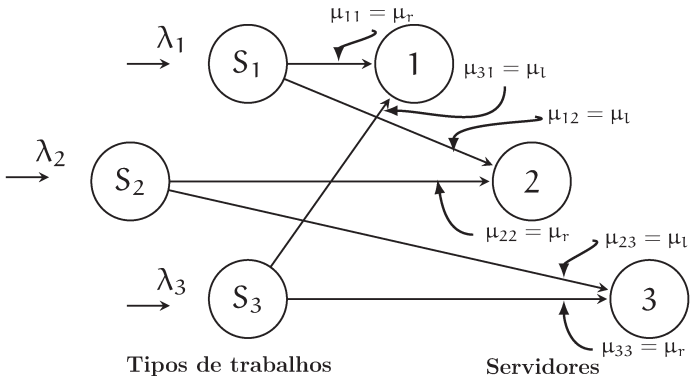


Figura 3.1: Representação do exemplo de rede.

Desta forma, para cada trabalho i que chega independentemente à rede, de acordo com um processo de Poisson $\lambda_i > 0$, $i = 1, 2, 3$, usando alguma política previamente estabelecida, são encaminhados a um servidor (ou fila) com a taxa de atendimento lenta ou rápida para serem processados.

No entanto, observe que se as taxas dos atendimentos forem trocadas para um único tipo de trabalho do exemplo de rede fornecido pela Figura 3.1, ou seja,

$$\mu_{1j} = \begin{cases} \mu_r & \text{se j=i} \\ \mu_l & \text{se j}\neq i \end{cases}$$

$$\mu_{2j} = \begin{cases} \mu_r & \text{se j=i} \\ \mu_l & \text{se j}\neq i \end{cases}$$

$$\mu_{3j} = \begin{cases} \mu_l & \text{se j=i} \\ \mu_r & \text{se j}\neq i \end{cases},$$

(veja a Figura 3.2) esta rede se tornará instável (ou ineficiente) para as políticas estáticas JFQ.

Isto acontecerá porque os trabalhos vão ser encaminhados para os servidores com a taxa mais rápida de atendimento e sobrecarregará o servidor 1.

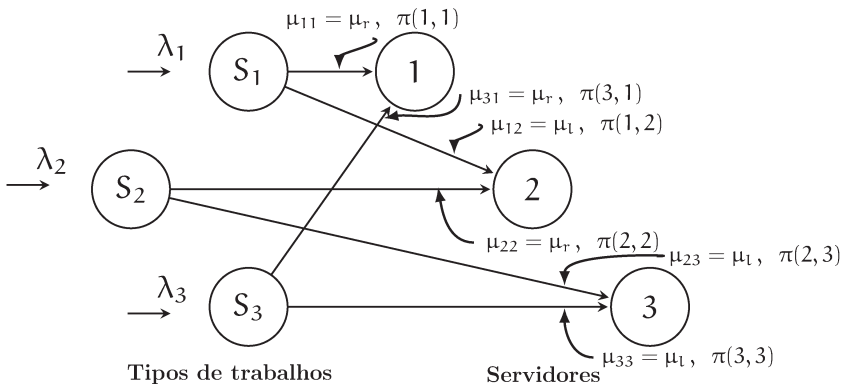


Figura 3.2: Representação do contraexemplo da rede G.

Contudo, é possível transformar esta rede de filas instável numa rede estável, desde que as probabilidades dos encaminhamentos dos trabalhos aos servidores sejam adequadas, ou seja, desde que as probabilidades satisfaçam a igualdade (3.1) e a restrição (3.2).

$$\frac{\pi(1,1)}{\mu_r} + \frac{1-\pi(3,3)}{\mu_r} = \frac{1-\pi(1,1)}{\mu_l} + \frac{\pi(2,2)}{\mu_r} = \frac{1-\pi(2,2)}{\mu_l} + \frac{\pi(3,3)}{\mu_l} \tag{3.1}$$

$$\min_{\pi(2,2)\in[\mathcal{R};1]} \left\{ \left(\frac{\pi(1,1)}{\mu_r} + \frac{1-\pi(3,3)}{\mu_r} \right) = \left(\frac{1-\pi(2,2)}{\mu_l} + \frac{\pi(3,3)}{\mu_l} \right) \right\} \tag{3.2}$$

$$\text{onde } \mathcal{R} = \frac{\mu_r^2 - \mu_l\mu_r}{\mu_r\mu_l + \mu_r^2 + \mu_l^2}$$

Isso significa que a rede fornecida pela Figura 3.2 será estável se as probabilidades dos encaminhamentos dos trabalhos aos servidores da rede, satisfazendo a igualdade (3.1) e a restrição (3.2), são

$$\pi(2,2) \in [\mathcal{R};1] \text{ , onde } \mathcal{R} = \frac{\mu_r^2 - \mu_l\mu_r}{\mu_r\mu_l + \mu_r^2 + \mu_l^2} \text{ ;}$$
$$\pi(2,3) = 1 - \pi(2,2) \text{ ;}$$

$$\pi(1,2) = 1 - \pi(1,1)\,;$$
$$\pi(3,1) = 1 - \pi(3,3)\,;$$
$$\pi(1,1) = \frac{\mu_r^3 - \mu_r\mu_l^2 + \left(3\mu_r\mu_l^2 + 2\mu_r^2\mu_l + \mu_l^3\right)\pi(2,2)}{3\mu_r^2\mu_l + 2\mu_r\mu_l^2 + \mu_r^3} \text{ e}$$
$$\pi(3,3) = \frac{\mu_r\mu_l - \mu_r^2 + \left(\mu_l^2 + \mu_r\mu_l + \mu_r^2\right)\pi(2,2)}{2\mu_r\mu_l + \mu_r^2}.$$

Com o objetivo de analisar de modo eficiente as comparações entre as políticas estáticas e dinâmicas, este trabalho considerou somente as políticas estáticas que apresentam as probabilidades adequadas dos encaminhamentos dos trabalhos aos servidores, garantindo a estabilidade da rede. As políticas de escolha de servidor com estas características são denominadas de melhor política estática JSW_b.

Desta forma, os exemplos de redes apresentados pelas Figuras 3.1 e 3.2 foram simulados utilizando as três políticas de escolha previamente estabelecidas, ou seja, as políticas JFQ_b, JSWA e JSWAS, durante um tempo fixo $T = 18000$ minutos, ou seja, 300 horas, cujos trabalhos i chegam independentemente à rede com taxa $\lambda_i = 1$, $i = 1,2,3$, podendo ser processados, dependendo do servidor escolhido, com taxas μ_r=1,2 ou μ_l=0,7.

3.2 Algoritmos das políticas de escolha

A ideia principal dos algoritmos dos encaminhamentos dos trabalhos aos servidores da rede para cada uma das políticas de escolha para os processos Y, X e ξ é apresentada nesta seção. Em seguida, os resultados obtidos e as conclusões das simulações que fornecem a política ideal de escolha de servidor para os dois exemplos de redes de filas são exibidos.

Lembre-se de que na construção do processo X quando existe a carga horária média total de trabalho $x_j < 0$ é definido $x_j \mapsto 0$, para todo $j \in N$. Isso significa que os servidores só processam os trabalhos que estão nas suas filas e não realizam nenhum trabalho em seu tempo ocioso.

Enquanto que para o processo ξ, nas situações em que existem a carga horária média total de trabalho $x_j < 0$, para todo servidor $j \in N$, significa que existe uma grande quantidade de pedidos de um determinado tipo de produto e que as empresas

precisam liberar, mas devido a indisponibilidade do produto exigido, com o tempo estas empresas acabam ficando com os seus pedidos em pendência no sistema.

No entanto, para comparar os desempenhos dos dois exemplos de redes de filas de acordo com o processo Y com os desempenhos obtidos pelos processos X e ξ, é necessário analisar o processo Y de acordo com a velocidade média estimada em relação à quantidade (ou número) de trabalhos e em relação ao tempo médio dos atendimentos para cada um dos servidores da rede. Note que a comparação deve ser realizada desta forma porque os processos ξ e X analisam o desempenho das redes de acordo a velocidade média estimada em relação à carga horária média total de trabalho dos servidores da rede.

Portanto, para todo estado

$$\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_j}, \ldots, \mathbf{y_N})$$

que apresenta a configuração atual dos trabalhos na fila dos N servidores da rede, onde

$$\mathbf{y_j} = (\mathcal{Y}_1, \ldots, \mathcal{Y}_l)$$

é o vetor que apresenta a sequência dos tipos de trabalho $\mathcal{Y}_l \in \{1, \ldots, \mathcal{K}\}$ que estão na fila do servidor j na posição l da configuração atual da rede, cuja taxa de atendimento será $\mu_{\mathcal{Y}_l j}$.

Então, a velocidade média em relação à quantidade (ou número) de trabalhos nos servidores é dada por

$$V_N(j, \mathbf{y}, \pi) = |y_j|/T$$

E a velocidade média em relação à carga horária média total de trabalho nos servidores é dada por

$$V_T(j, \mathbf{y}, \pi) = (\sum_{l=1}^{|y_j|} 1/\mu_{\mathcal{Y}_l j})/T$$

Assim, os processos ξ e X apresentarão um comportamento semelhante ao processo Y se

$$V_T(j, \mathbf{y}, \pi) = (\sum_{l=1}^{|y_j|} 1/\mu_{\mathcal{Y}_l j})/T \;\approx\; V(j, \xi, \pi) \;\approx\; V(j, \mathbf{x}, \pi)$$

Algoritmo da melhor política estática

Para a melhor política estática de escolha de servidor JFQ_b, os trabalhos i que chegam à rede são encaminhados para os servidores de acordo com as probabilidades $\pi(i,j)$ adequadas, para todo $i = 1, 2$ e 3; e $j = 1, 2$ e 3.

As entradas e saídas dos tipos de trabalho são distribuídos uniformemente de acordo com as taxas das chegadas e dos atendimentos.

A 1. Ideia do algoritmo para a política JFQ_b segundo o processo Y

Início do algoritmo

1. Enquanto $t \leq T$,

   lugar.j $\leftarrow 0$

   qtos.clientes.j $\leftarrow 0$

   saida.j $\leftarrow 0$

   carga.horaria.j $\leftarrow 0$

   (a) se um cliente de trabalho i chegou à rede, então

      i. o trabalho é encaminhado com probabilidade adequada $\pi(j,j)$ para o servidor j que processará o trabalho com a taxa de serviço μ_{ij}.
      Em seguida, guarda-se a taxa do atendimento na posição lugar.j do vetor Taxa.serviço.j(lugar.j), indicando a posição do cliente na fila e a taxa que o servidor processará o trabalho.

$$\begin{aligned} \text{lugar.j} &\leftarrow \text{lugar.j} + 1 \\ \text{Taxa.serviço.j(lugar.j)} &\leftarrow \mu_{ij} \\ \text{qtos.clientes.j} &\leftarrow \text{qtos.clientes.j} + 1 \\ \text{carga.horaria.j} &\leftarrow \text{carga.horaria.j} + 1/\mu_{ij} \end{aligned}$$

   (b) se o servidor j processou por completo o trabalho do cliente que estava na posição saida.j com taxa Taxa.serviço.j(saida.j). Então,

$$\begin{aligned} \text{saida.j} &\leftarrow \text{saida.j} + 1 \\ \text{qtos.clientes.j} &\leftarrow \text{qtos.clientes.j} - 1 \\ \text{carga.horaria.j} &\leftarrow \text{carga.horaria.j} - 1/\text{Taxa.serviço.j(saida.j)} \end{aligned}$$

2. Se $t \geq T$,

   Calcula-se a velocidade média em relação a quantidade (ou número) de trabalhos e a velocidade média em relação à carga horária média total de trabalho do servidor j, para todo $j = 1, 2$ e 3.

$$V_N(j, y, JFQ) \leftarrow \text{qtos.clientes.j } /t$$
$$V_T(j, y, JFQ) \leftarrow \text{carga.horaria.j } /t$$

Fim do algoritmo

A 2. Ideia do algoritmo para a política JFQ_b segundo o processo ξ

Início do algoritmo

1. Enquanto $t \leq T$,
   $\xi_j \leftarrow 0$

   (a) se um cliente de trabalho i chegou à rede, então

      i. o trabalho é encaminhado com probabilidade adequada $\pi(j, j)$ para o servidor j que processará o trabalho com a taxa de serviço μ_{ij}. Em seguida, a carga horária média total de trabalho do servidor j é atualizada

$$\xi_j \leftarrow \xi_j + (1/\mu_{ij})$$

   (b) se o servidor j processou por completo, em média, um trabalho com carga horária média total de trabalho $(1/\mu_{ij})$. Então, a carga horária média total de trabalho deste servidor é atualizada

$$\xi_j \leftarrow \xi_j - 1$$

2. Se $t \geq T$,

   Calcula-se a velocidade média estimada para a carga horária média total de trabalho $V(j, \xi, JFQ_b)$ do servidor j, para todo $j = 1, 2$ e 3,

$$V(j, \xi, JFQ_b) \leftarrow \xi_j/t$$

Fim do algoritmo

A 3. Ideia do algoritmo para a política JFQ_b segundo o processo X

Início do algoritmo

1. Enquanto $t \leq T$,
   $x_j \leftarrow 0$

   (a) se um cliente de trabalho i chegou à rede, então

      i. o trabalho é encaminhado com probabilidade adequada $\pi(j, j)$ para o servidor j que processará o trabalho com a taxa de serviço μ_{ij}.
      Em seguida, a carga horária média total de trabalho do servidor j é atualizada

$$x_j \leftarrow x_j + (1/\mu_{ij})$$

   (b) se o servidor j processou por completo, em média, um trabalho com carga horária média total de trabalho $(1/\mu_{ij})$. Então, a carga horária média total de trabalho deste servidor é atualizada

$$x_j \leftarrow x_j - 1$$

$$\text{se } x_j < 0\,, \text{ então } x_j \leftarrow 0$$

2. Se $t \geq T$,
   Calcula-se a velocidade média estimada para a carga horária média total de trabalho $V(j, x, JFQ_b)$ do servidor j, para todo $j = 1, 2$ e 3,

$$V(j, x, JFQ_b) \leftarrow x_j/t$$

Fim do algoritmo

Considerando o contraexemplo apresentado pela Figura 3.2 e os processos X e ξ, as probabilidades dos encaminhamentos dos trabalhos aos servidores de acordo com as taxas simuladas, ou seja, para as taxas dos atendimentos, $\mu_r = 1,2$ e $\mu_l = 0,7$, satisfazendo

$$\frac{\pi(1,1)}{\mu_r} + \frac{1-\pi(3,3)}{\mu_r} = \frac{1-\pi(1,1)}{\mu_l} + \frac{\pi(2,2)}{\mu_r} = \frac{1-\pi(2,2)}{\mu_l} + \frac{\pi(3,3)}{\mu_l} \text{ e}$$

$$\min_{\pi(2,2)\in(0{,}29;1]} \left\{ \left(\frac{\pi(1,1)}{\mu_r} + \frac{1-\pi(3,3)}{\mu_r} \right) = \left(\frac{1-\pi(2,2)}{\mu_l} + \frac{\pi(3,3)}{\mu_l} \right) \right\} \text{ são as seguintes:}$$

$$\begin{aligned} \pi(1,1) &= 0,89 \quad & \pi(1,2) &= 0,11 \\ \pi(2,2) &= 1,00 \quad & \pi(2,3) &= 0,00 \\ \pi(3,3) &= 0,69 \quad & \pi(3,1) &= 0,31 \end{aligned}$$

É importante ressaltar que as probabilidades dos encaminhamentos dos trabalhos aos servidores do contraexemplo para os processos Y, X e ξ, são as mesmas, a fim de comparar os processos de forma eficiente.

Ideia do algoritmo para as políticas dinâmicas

Para a análise das políticas dinâmicas foram considerados dois cenários para os encaminhados dos trabalhos aos servidores:

1. a política que encaminha os trabalhos de acordo com a menor carga horária média total de trabalho da fila, ponderada pelo tempo médio do processamento dos servidores escolhidos JSWA; e
2. a política que encaminha os trabalhos de acordo com a menor carga horária média total de trabalho do sistema, ponderada pelo tempo médio do processamento dos servidores escolhidos JSWAS.

Entretanto, se os servidores apresentam a mesma carga horária média total de trabalho ponderada pelo tempo médio do processamento do servidor escolhido, então os trabalhos que chegam à rede juntam-se à fila do servidor que possui a taxa de atendimento mais rápida μ_r, o que é mais conveniente.

B 1. Ideia do algoritmo para a política JSWA segundo o processo Y

Início do algoritmo

1. Enquanto $t \leq T$,

   qtos.clientes.j $\leftarrow 0$

   carga.horaria.j $\leftarrow 0$

   lugar.j $\leftarrow 0$

   saida.j $\leftarrow 0$

   (a) se um cliente de trabalho i chegou à rede, então

   i. se $\left((\sum_{l=1}^{|y_j|} 1/\mu_{y_l j})/\mu_{ij} < (\sum_{l=1}^{|y_r|} 1/\mu_{y_l r})/\mu_{ir} \right)$, ele é encaminhado para o servidor j que realizará o processamento com taxa de serviço μ_{ij}. Em seguida, guarda-se a taxa do atendimento na posição lugar.j do vetor Taxa.serviço.j(lugar.j), indicando a posição do cliente na fila e a taxa que o servidor processará o trabalho.

   $$\begin{aligned} \text{lugar.j} &\leftarrow \text{lugar.j} + 1 \\ \text{Taxa.serviço.j(lugar.j)} &\leftarrow \mu_{ij} \\ \text{qtos.clientes.j} &\leftarrow \text{qtos.clientes.j} + 1 \\ \text{carga.horaria.j} &\leftarrow \text{carga.horaria.j} + 1/\mu_{ij} \end{aligned}$$

   ii. se $\left((\sum_{l=1}^{|y_j|} 1/\mu_{y_l j})/\mu_{ij} = (\sum_{l=1}^{|y_r|} 1/\mu_{y_l r})/\mu_{ir} \right)$, ele é encaminhado ao servidor $j \in \tilde{\mathcal{B}}_i(y)$ tal que $\mu_{ij} = \max_{l \in \tilde{\mathcal{B}}_i(y)} \mu_{il}$, onde

   $$\tilde{\mathcal{B}}_i(y) = \left\{ j, r \in S_i : (\sum_{l=1}^{|y_j|} 1/\mu_{y_l j})/\mu_{ij} = (\sum_{l=1}^{|y_r|} 1/\mu_{y_l r})/\mu_{ir} \right\}.$$

   Em seguida, guarda-se a taxa do atendimento na posição lugar.j do vetor Taxa.serviço.j(lugar.j), indicando a posição do cliente na fila e a taxa que o servidor processará o trabalho.

   $$\begin{aligned} \text{lugar.j} &\leftarrow \text{lugar.j} + 1 \\ \text{Taxa.serviço.j(lugar.j)} &\leftarrow \mu_{ij} \\ \text{qtos.clientes.j} &\leftarrow \text{qtos.clientes.j} + 1 \\ \text{carga.horaria.j} &\leftarrow \text{carga.horaria.j} + 1/\mu_{ij} \end{aligned}$$

(b) se o servidor j processou por completo o trabalho do cliente que estava na posição saida.j com taxa Taxa.serviço.j(saida.j). Então,

$$\text{saida.j} \leftarrow \text{saida.j} + 1$$
$$\text{qtos.clientes.j} \leftarrow \text{qtos.clientes.j} - 1$$
$$\text{carga.horaria.j} \leftarrow \text{carga.horaria.j} - 1/\text{Taxa.serviço.j(saida.j)}$$

2. Se $t \geq T$,

Calcula-se a velocidade média em relação à quantidade (ou número) de trabalhos e a velocidade média em relação à carga horária média total de trabalho do servidor j, para todo $j = 1, 2$ e 3.

$$V_N(j, y, \text{JSWA}) \leftarrow \text{qtos.clientes.j } /t$$

$$V_T(j, y, \text{JSWA}) \leftarrow \text{carga.horaria.j } /t$$

Fim do algoritmo

B 2. Ideia do algoritmo para a política JSWA segundo o processo ξ

Início do algoritmo

1. Enquanto $t \leq T$,

$\xi_j(t) \leftarrow 0$

$\xi_l(t) \leftarrow 0$

(a) se um trabalho i chegou à rede, então

i. se $(\xi_j(t)/\mu_{ij} < \xi_l(t)/\mu_{il})$, ele é encaminhado para o servidor j que realizará o processamento com taxa de serviço μ_{ij}. Em seguida, a carga horária média total de trabalho do servidor j é atualizada

$$\xi_j(t) \leftarrow \xi_j(t) + (1/\mu_{ij})$$

ii. se $(\xi_j(t)/\mu_{ij} = \xi_l(t)/\mu_{il})$, para todo $(j, l) \in S_i$, ele é encaminhado ao servidor $j \in \tilde{\mathcal{B}}_i(\xi)$ tal que $\mu_{ij} = \max_{l \in \tilde{\mathcal{B}}_i(\xi)} \mu_{il}$, onde

$$\tilde{\mathcal{B}}_i(\xi) = \{j, l \in S_i : \xi_j/\mu_{ij} = \xi_l/\mu_{il}\}.$$

Em seguida, a carga horária média total de trabalho do servidor j é atualizada

$$\xi_j(t) \leftarrow \xi_j(t) + (1/\mu_{ij})$$

(b) se o servidor j processou por completo, em média, um trabalho com carga horária média total de trabalho $(1/\mu_{ij})$. Então, realiza-se a atualização com taxa 1,

$$\xi_j(t) \leftarrow \xi_j(t) - 1$$

2. Se $t \geq T$,

Calcula-se a velocidade média estimada da carga horária média total de trabalho $V(j, \xi, \text{JSWA})$ para todo $j = 1, 2$ e 3, dada por

$$V(j, \xi, \text{JSWA}) \leftarrow \xi_j(t)/t$$

Fim do algoritmo

B 3. Ideia do algoritmo para a política JSWA segundo o processo X

Início do algoritmo

1. Enquanto $t \leq T$,

$x_j(t) \leftarrow 0$

$x_l(t) \leftarrow 0$

(a) se um trabalho i chegou à rede, então

i. se $(x_j(t)/\mu_{ij} < x_l(t)/\mu_{il})$, ele é encaminhado para o servidor j que realizará o processamento com taxa de serviço μ_{ij}. Em seguida, a carga horária média total de trabalho do servidor j é atualizada

$$x_j(t) \leftarrow x_j(t) + (1/\mu_{ij})$$

ii. se $(x_j(t)/\mu_{ij} = x_l(t)/\mu_{il})$, para todo $(j, l) \in S_i$, ele é encaminhado ao servidor $j \in \tilde{\mathcal{B}}_i(x)$ tal que $\mu_{ij} = \max_{l \in \tilde{\mathcal{B}}_i(x)} \mu_{il}$, onde

$$\tilde{\mathcal{B}}_i(x) = \{j, l \in S_i : x_j/\mu_{ij} = x_l/\mu_{il}\}\,.$$

Em seguida, a carga horária média total de trabalho do servidor j é atualizada

$$x_j(t) \leftarrow x_j(t) + (1/\mu_{ij})$$

(b) se o servidor j processou por completo, em média, um trabalho com carga horária média total de trabalho $(1/\mu_{ij})$. Então, realiza-se a atualização com taxa 1,

$$x_j(t) \leftarrow x_j(t) - 1$$

$$\text{se } x_j(t) < 0\,, \text{ então } x_j(t) \leftarrow 0$$

2. Se $t \geq T$,

Calcula-se a velocidade média estimada da carga horária média total de trabalho $V(j, x, \text{JSWA})$ para todo $j = 1, 2$ e 3, dada por

$$V(j, x, \text{JSWA}) \leftarrow x_j(t)/t$$

Fim do algoritmo

C 1. Ideia do algoritmo para a política JSWAS segundo o processo Y

Início do algoritmo

1. Enquanto $t \leq T$,

qtos.clientes.j $\leftarrow 0$

carga.horaria.j $\leftarrow 0$

lugar.j $\leftarrow 0$

saida.j $\leftarrow 0$

(a) se um cliente de trabalho i chegou à rede, então

i. se $\left[\left((\sum_{l=1}^{|y_j|} 1/\mu_{y_l j}) + 1/\mu_{ij} \right) /\mu_{ij} < \left((\sum_{l=1}^{|y_r|} 1/\mu_{y_l r}) + 1/\mu_{ir} \right) /\mu_{ir} \right]$, ele é encaminhado para o servidor j que realizará o processamento com taxa de serviço μ_{ij}. Em seguida, guarda-se a taxa do atendimento na posição lugar.j do vetor Taxa.serviço.j(lugar.j), indicando a posição do cliente na fila e a taxa que o servidor processará o trabalho.

$$\text{lugar.j} \leftarrow \text{lugar.j} + 1$$

$$\text{Taxa.serviço.j(lugar.j)} \leftarrow \mu_{ij}$$
$$\text{qtos.clientes.j} \leftarrow \text{qtos.clientes.j} + 1$$
$$\text{carga.horaria.j} \leftarrow \text{carga.horaria.j} + 1/\mu_{ij}$$

ii. se $\left[\left((\sum_{l=1}^{|y_j|} 1/\mu_{y_l j}) + 1/\mu_{ij} \right) /\mu_{ij} < \left((\sum_{l=1}^{|y_r|} 1/\mu_{y_l r}) + 1/\mu_{ir} \right) /\mu_{ir} \right]$, ele é encaminhado ao servidor $j \in \tilde{\mathcal{B}}_i(y)$ tal que $\mu_{ij} = \max_{l \in \tilde{\mathcal{B}}_i(y)} \mu_{il}$, onde

$$\tilde{\mathcal{B}}_i(y) = \left\{ j, r \in S_i : \left((\sum_{l=1}^{|y_j|} 1/\mu_{y_l j}) + 1/\mu_{ij} \right) /\mu_{ij} = \left((\sum_{l=1}^{|y_r|} 1/\mu_{y_l r}) + 1/\mu_{ir} \right) /\mu_{ir} \right\}.$$

Em seguida, guarda-se a taxa do atendimento na posição lugar.j do vetor Taxa.serviço.j(lugar.j), indicando a posição do cliente na fila e a taxa que o servidor processará o trabalho.

$$\text{lugar.j} \leftarrow \text{lugar.j} + 1$$
$$\text{Taxa.serviço.j(lugar.j)} \leftarrow \mu_{ij}$$
$$\text{qtos.clientes.j} \leftarrow \text{qtos.clientes.j} + 1$$
$$\text{carga.horaria.j} \leftarrow \text{carga.horaria.j} + 1/\mu_{ij}$$

(b) se o servidor j processou por completo o trabalho do cliente que estava na posição saida.j com taxa Taxa.serviço.j(saida.j). Então,

$$\text{saida.j} \leftarrow \text{saida.j} + 1$$
$$\text{qtos.clientes.j} \leftarrow \text{qtos.clientes.j} - 1$$
$$\text{carga.horaria.j} \leftarrow \text{carga.horaria.j} - 1/\text{Taxa.serviço.j(saida.j)}$$

2. Se $t \geq T$,

Calcula-se a velocidade média em relação à quantidade (ou número) de trabalhos e a velocidade média em relação à carga horária média total de trabalho do servidor j, para todo $j = 1, 2$ e 3.

$$V_N(j, y, \text{JSWAS}) \leftarrow \text{qtos.clientes.j } /t$$
$$V_T(j, y, \text{JSWAS}) \leftarrow \text{carga.horaria.j } /t$$

Fim do algoritmo

C 2. Ideia do algoritmo para a política JSWAS segundo o processo ξ

Início do algoritmo

1. Enquanto $t \leq T$,
   $\xi_j(t) \leftarrow 0$
   $\xi_l(t) \leftarrow 0$

   (a) se um trabalho i chegou à rede, então

      i. se $\left[\left(\xi_j(t) + 1/\mu_{ij}\right)/\mu_{ij} \; < \; \left(\xi_l(t) + 1/\mu_{il}\right)/\mu_{il} \right]$, ele é encaminhado para o servidor j que realizará o processamento com a taxa de serviço μ_{ij}. Em seguida, a carga horária média total de trabalho do servidor j é atualizada

$$\xi_j(t) \leftarrow \xi_j(t) + (1/\mu_{ij})$$

      ii. se $\left[\left(\xi_j(t) + 1/\mu_{ij}\right)/\mu_{ij} = \left(\xi_l(t) + 1/\mu_{il}\right)/\mu_{il} \right]$ para todo $(j, l) \in S_i$, ele é encaminhado para o servidor $j \in \tilde{\mathcal{B}}_i(\xi)$ tal que $\mu_{ij} = \max\limits_{l \in \tilde{\mathcal{B}}_i(\xi)} \mu_{il}$, onde

$$\tilde{\mathcal{B}}_i(\xi) = \left\{ j, l \in S_i : \left[\left(\xi_j(t) + 1/\mu_{ij}\right)/\mu_{ij} = \left(\xi_l(t) + 1/\mu_{il}\right)/\mu_{il} \right] \right\} .$$

      Em seguida, a carga horária média total de trabalho do servidor j é atualizada

$$\xi_j(t) \leftarrow \xi_j(t) + (1/\mu_{ij})$$

   (b) se o servidor j processou por completo, em média, um trabalho com carga horária média total de trabalho $(1/\mu_{ij})$. Então, realiza-se a atualização com taxa 1,

$$\xi_j(t) \leftarrow \xi_j(t) - 1$$

2. Se $t \geq T$,

   Calcula-se a velocidade média estimada para a carga horária média total de trabalho $V(j, \xi, \text{JSWAS})$ para todo $j = 1, 2$ e 3, dada por

$$V(j, \xi, \text{JSWAS}) \leftarrow \xi_j(t)/t$$

Fim do algoritmo

C 3. Ideia do algoritmo para a política JSWAS segundo o processo X

Início do algoritmo

1. Enquanto $t \leq T$,

   $x_j(t) \leftarrow 0$

   $x_l(t) \leftarrow 0$

   (a) se um trabalho i chegou à rede, então

   i. se $\left[(x_j(t) + 1/\mu_{ij})/\mu_{ij} < (x_l(t) + 1/\mu_{il})/\mu_{il} \right]$, ele é encaminhado para o servidor j que realizará o processamento com a taxa de serviço μ_{ij}. Em seguida, a carga horária média total de trabalho do servidor j é atualizada

   $$x_j(t) \leftarrow x_j(t) + (1/\mu_{ij})$$

   ii. se $\left[(x_j(t) + 1/\mu_{ij})/\mu_{ij} = (x_l(t) + 1/\mu_{il})/\mu_{il} \right]$ para todo $(j, l) \in S_i$, ele é encaminhado para o servidor $j \in \tilde{\mathcal{B}}_i(\xi)$ tal que $\mu_{ij} = \max_{l \in \tilde{\mathcal{B}}_i(x)} \mu_{il}$, onde

   $$\tilde{\mathcal{B}}_i(x) = \left\{ j, l \in S_i : \left[(x_j(t) + 1/\mu_{ij})/\mu_{ij} = (x_l(t) + 1/\mu_{il})/\mu_{il} \right] \right\} .$$

   Em seguida, a carga horária média total de trabalho do servidor j é atualizada

   $$x_j(t) \leftarrow x_j(t) + (1/\mu_{ij})$$

   (b) se o servidor j processou por completo, em média, um trabalho com carga horária média total de trabalho $(1/\mu_{ij})$. Então, realiza-se a atualização com taxa 1,

   $$x_j(t) \leftarrow x_j(t) - 1$$

   $$\text{se } x_j(t) < 0\,, \text{ então } x_j(t) \leftarrow 0$$

2. Se $t \geq T$,

   Calcula-se a velocidade média estimada para a carga horária média total

de trabalho $V(j, x, \text{JSWAS})$ para todo $j = 1, 2$ e 3, dada por

$$V(j, x, \text{JSWAS}) \leftarrow x_j(t)/t$$

Fim do algoritmo

3.3 Resultados e conclusões das simulações

Nesta seção são apresentados os resultados e as conclusões obtidas das simulações para os dois exemplos de redes de filas apresentados pelas Figuras 3.1 e 3.2 de acordo com os processos Y, ξ e X, segundo as políticas de escolha JFQ_b, JSWA e JSWAS.

Resultados das simulações e conclusões para o processo Y

Para comparar os desempenhos dos dois exemplos de redes de filas de acordo com o processo Y com os desempenhos obtidos pelos processos X e ξ, é necessário analisar o processo Y de acordo com a velocidade média estimada da fila em relação à quantidade (ou número) de trabalho que estão nos servidores e de acordo com a carga horária média total de trabalho de cada um dos servidores da rede.

Note que a comparação do processo Y deve ser realizada desta forma porque os processos ξ e X analisam o desempenho das redes de acordo a velocidade média estimada da carga horária média total de trabalho nos servidores da rede.

Desta forma, a Tabela 3.1 apresenta a velocidade média estimada em relação ao número de trabalhos nos servidores do exemplo de rede da Figura 3.1 para o processo Y, durante um tempo fixo $T = 18000$ minutos, ou seja, 300 horas.

Enquanto que a Tabela 3.2 apresenta a velocidade média estimada em relação à carga horária média total de trabalho dos servidores para este exemplo de rede.

Tabela 3.1: Velocidade média estimada em relação ao número de trabalhos para os servidores do exemplo de rede de filas dado pela Figura 3.1 segundo o processo Y

	JFQ_b	JSWA	JSWAS
Servidor 1	0.000222	0.000000	0.000007
Servidor 2	0.000222	0.000000	0.000000
Servidor 3	0.000000	0.109555	0.1072776

Tabela 3.2: Velocidade média estimada em relação à carga horária média total de trabalho para os servidores do exemplo de rede de filas dado pela Figura 3.1 segundo o processo Y

	JFQ_b	JSWA	JSWAS
Servidor 1	0.000185	0.000000	0.000046
Servidor 2	0.000185	0.000000	0.000000
Servidor 3	0.000000	0.112295	0.110397

A velocidade nula de alguns servidores significa que não existe nenhum trabalho na fila destes servidores e que eles estão aguardando algum trabalho chegar, ou seja, todos os trabalhos que chegaram já foram completamente processados.

Para o contraexemplo de rede apresentado pela Figura 3.2, a Tabela 3.3 exibe a velocidade média estimada em relação ao número de trabalhos nos servidores e a Tabela 3.4 apresenta a velocidade média estimada em relação à carga horária média total de trabalho nos servidores, durante um tempo fixo $T = 18000$ minutos, ou seja, 300 horas.

Tabela 3.3: Velocidade média estimada em relação ao número de trabalhos para os servidores do contraexemplo de rede de filas dado pela Figura 3.2 segundo o processo Y

	JFQ_b	JSWA	JSWAS
Servidor 1	0.022444	0.208609	0.206384
Servidor 2	0.003611	0.000111	0.000000
Servidor 3	0.001499	0.071110	0.070498

Tabela 3.4: Velocidade média estimada em relação à carga horária média total para os servidores do contraexemplo de rede de filas dado pela Figura 3.2 segundo o processo Y

	JFQ_b	JSWA	JSWAS
Servidor 1	0.018703	0.1738405	0.171986
Servidor 2	0.003108	0.0000952	0.000000
Servidor 3	0.002143	0.1015861	0.100712

Observe que a velocidade média estimada da carga horária média total de trabalho dos servidores do processo Y apresentada acima pelas Tabelas 3.2 e 3.4 pode ser hierarquizada de acordo com os agrupamentos de servidores como estabelecem as Tabelas 3.5 e 3.6.

Tabela 3.5: Hierarquização da velocidade média estimada da carga horária média total de trabalho dos servidores do exemplo de rede de filas dado pela Figura 3.1 de acordo com o processo Y

Política	Agrupamento	Servidor	Velocidade
JFQ_b	C_1	1, 2 e 3	≈ 0.00
JSWA	C_1	3	≈ 0.11
	C_2	1 e 2	≈ 0.00
JSWAS	C_1	3	≈ 0.11
	C_2	1 e 2	≈ 0.00

Tabela 3.6: Hierarquização da velocidade média estimada da carga horária média total de trabalho dos servidores do contraexemplo de rede de filas dado pela Figura 3.2 de acordo com o processo Y

Política	Agrupamento	Servidor	Velocidade
JFQ_b	C_1	1	≈ 0.02
	C_2	2 e 3	≈ 0.00
JSWA	C_1	1	≈ 0.17
	C_2	2	≈ 0.10
	C_3	3	≈ 0.00
JSWAS	C_1	1	≈ 0.17
	C_2	2	≈ 0.10
	C_3	3	≈ 0.00

Lembre-se de que a política ideal de escolha de servidor é obtida através da análise da velocidade média estimada dos servidores (ou das filas) dentro dos agrupamentos C de servidores submetidos a um conjunto de políticas previamente estabelecidas.

Portanto, a política ideal de escolha de servidor é aquela política que oferece

$$\mathcal{V}(C; \mu_{ij}, \Pi') = \min_{\pi \in \Pi'} \max_{j \in C} V(j; \pi) \ ,$$

ou seja, o mínimo (sobre todas as políticas Π') da velocidade máxima da fila dos servidores que estão dentro do agrupamento C.

Então, considerando as Tabelas 3.5 e 3.6 , percebe-se que a política JSWA oferece um desempenho semelhante em relação ao desempenho fornecido pela política JSWAS. No entanto, estas políticas dinâmicas (JSWA e JSWAS) não oferecem um

desempenho superior ao encontrado pela melhor política estática $\mathrm{JFQ_b}$.

Portanto, é possível concluir que a política ideal de escolha de servidor para o processo Y é a melhor política estática.

Resultados das simulações e conclusões para o processo ξ

Para o processo ξ, a Tabela 3.7 apresenta a velocidade média estimada da carga horária média total de trabalho para os servidores do exemplo de rede de filas apresentado pela Figura 3.1, durante um tempo fixo de $\mathsf{T} = 18000$ minutos, ou seja, 300 horas.

Enquanto que a Tabela 3.8 exibe os resultados obtidos para o contraexemplo de rede de filas apresentado pela Figura 3.2.

Tabela 3.7: Velocidade média estimada da carga horária média total de trabalho dos servidores do exemplo de rede de filas dado pela Figura 3.1 de acordo com o processo ξ

	$\mathrm{JFQ_b}$	JSWA	JSWAS
Servidor 1	-0.1720553	0.0009735432	0.0008412684
Servidor 2	-0.1735830	0.0005648138	0.0004325389
Servidor 3	-0.1684441	0.0007817446	0.0006494697

Tabela 3.8: Velocidade média estimada da carga horária média total de trabalho dos servidores do contraexemplo de rede de filas dado pela Figura 3.2 de acordo com o processo ξ

	$\mathrm{JFQ_b}$	JSWA	JSWAS
Servidor 1	-0.004913	0.001953694	0.001768510
Servidor 2	-0.007587	0.001458988	0.001326713
Servidor 3	-0.031727	0.000714282	0.000634918

A velocidade negativa de alguns servidores do processo ξ significa que houve uma grande quantidade de pedidos de um determinado tipo de trabalho (ou produto) e que os servidores precisam liberar, mas por causa da indisponibilidade do trabalho (ou produto) exigido, com o tempo os servidores acabaram ficando com os seus trabalhos em pendência no sistema.

Observe novamente que a velocidade média estimada da carga horária média total de trabalho dos servidores do processo ξ, apresentada pelas Tabelas 3.7 e 3.8 pode ser hierarquizada de acordo com os agrupamentos de servidores como estabelecem as Tabelas 3.9 e 3.10.

Tabela 3.9: Hierarquização da velocidade média estimada da carga horária média total de trabalho dos servidores do exemplo de rede de filas dado pela Figura 3.1 de acordo com o processo ξ

Política	Agrupamento	Servidor	Velocidade
JFQ_b	C_1	1, 2 e 3	$\approx$ -0.17
JSWA	C_1	1, 2 e 3	$\approx$ 0.00
JSWAS	C_1	1, 2 e 3	$\approx$ 0.00

Tabela 3.10: Hierarquização da velocidade média estimada da carga horária média total de trabalho dos servidores do contraexemplo de rede de filas dado pela Figura 3.2 de acordo com o processo ξ

Política	Agrupamento	Servidor	Velocidade
JFQ_b	C_1	2	$\approx$ 0.00
	C_2	1	$\approx$ -0.01
	C_3	3	$\approx$ -0.03
JSWA	C_1	1, 2 e 3	$\approx$ 0.00
JSWAS	C_1	1, 2 e 3	$\approx$ 0.00

Relembre-se que a política ideal de escolha de servidor é aquela política que oferece

$$\mathcal{V}(C; \mu_{ij}, \Pi') = \min_{\pi \in \Pi'} \max_{j \in C} V(j; \pi) \ ,$$

ou seja, o mínimo (sobre todas as políticas Π') da velocidade máxima da fila dos servidores que estão dentro do agrupamento C de servidores.

Assim, usando as Tabelas 3.9 e 3.10, é possível perceber que a melhor política estática JFQ_b apresenta a velocidade dos servidores mais próxima em relação à velocidade teórica[1].

Portanto, a política JFQ_b oferece uma boa estabilidade para as redes com estas características de dependência dos atendimentos. Considerando somente as políticas dinâmicas, percebe-se que a política JSWAS fornece um desempenho igual ao obtido pela política JSWA. Entretanto, estas políticas dinâmicas não oferecem um desempenho superior ao obtido pela melhor política estática.

Então, a política ideal de escolha de servidor para processo ξ também é a melhor política estática.

Resultados das simulações e conclusões para o processo X

Para o processo X a Tabela 3.11 apresenta a velocidade média estimada da carga horária média total de trabalho dos servidores do exemplo de rede de filas apresentado pela Figura 3.1, durante um tempo fixo $T = 18000$ minutos, ou seja, 300 horas.

Enquanto que a Tabela 3.12 exibe a velocidade média estimada da carga horária média total de trabalho dos servidores do contraexemplo de rede de filas apresentado pela Figura 3.2.

[1] Cálculo da velocidade teórica

- para o exemplo de rede

$$V = \frac{1-\pi(2,2)}{\mu_l} + \frac{\pi(3,3)}{\mu_r} - 1 = 0,83 - 1,00 = -0,1666 = -0,17$$

- para o contraexemplo

$$V = \frac{1-\pi(2,2)}{\mu_l} + \frac{\pi(3,3)}{\mu_l} - 1 = 0,99 - 1,00 = -0,01$$

Tabela 3.11: Velocidade média estimada da carga horária média total de trabalho dos servidores do exemplo de rede de filas dado pela Figura 3.1 de acordo com o processo X

	$\mathrm{JFQ_b}$	JSWA	JSWAS
Servidor $\mathsf{1}$	0.000185	0.042227	0.038072
Servidor $\mathsf{2}$	0.000183	0.112080	0.112748
Servidor $\mathsf{3}$	0.000009	0.071262	0.065780

Tabela 3.12: Velocidade média estimada da carga horária média total de trabalho dos servidores do contraexemplo de rede de filas dado pela Figura 3.2 de acordo com o processo X

	$\mathrm{JFQ_b}$	JSWA	JSWAS
Servidor $\mathsf{1}$	0.018703	0.137230	0.140737
Servidor $\mathsf{2}$	0.009108	0.081494	0.085291
Servidor $\mathsf{3}$	0.002142	0.080633	0.085413

Note novamente que a velocidade média estimada da carga horária média total de trabalho dos servidores do processo X apresentada pelas Tabelas 3.11 e 3.12 pode ser hierarquizada de acordo com a velocidade dos agrupamentos de servidores como estabelecem as Tabelas 3.13 e 3.14.

Tabela 3.13: Hierarquização da velocidade média estimada da carga horária média total de trabalho dos servidores para o exemplo de rede de filas dado pela Figura 3.1 de acordo com processo X

Política	agrupamento	servidor	velocidade
$\mathrm{JFQ_b}$	C_1	$1, 2$ e 3	≈ 0.00
JSWA	C_1	2	≈ 0.11
	C_2	3	≈ 0.07
	C_3	1	≈ 0.04
JSWAS	C_1	2	≈ 0.11
	C_2	3	≈ 0.07
	C_3	1	≈ 0.04

Tabela 3.14: Hierarquização da velocidade média estimada da carga horária média total de trabalho dos servidores do contraexemplo de rede de filas dado pela Figura 3.2 de acordo com processo X

Política	Agrupamento	Servidor	Velocidade
JFQ_b	C_1	1	≈ 0.02
	C_2	2 e 3	≈ 0.00
JSWA	C_1	1	≈ 0.14
	C_2	2 e 3	≈ 0.08
JSWAS	C_1	1	≈ 0.14
	C_2	2 e 3	≈ 0.08

Utilizando-se novamente as Tabelas 3.13 e 3.14, procura-se a política ideal de escolha de servidor que apresenta

$$\mathcal{V}(C; \mu_{ij}, \Pi') = \min_{\pi \in \Pi'} \max_{j \in C} V(j; \pi) \ ,$$

ou seja, o mínimo (sobre todas as políticas Π') da velocidade máxima da fila dos servidores que estão dentro do agrupamento C de servidores.

Assim, é possível perceber que a melhor política estática JFQ_b oferece uma boa estabilidade para as redes com estas características de dependência dos atendimentos.

Considerando as políticas dinâmicas, nota-se novamente que a política JSWAS fornece um desempenho igual ao obtido pela política JSWA. Entretanto, estas políticas dinâmicas não oferecem um desempenho superior ao obtido pela melhor política estática JFQ_b.

Portanto, a política ideal de escolha de servidor para o processo X é a melhor política estática.

Conclusão geral das simulações

Foi possível perceber através das simulações que a melhor política estática JFQ_b é uma política ideal de escolha de servidor para os processos Y, X e ξ, de acordo com os exemplos simulados.

Além disso, o processo X é uma boa alternativa para analisar o comportamento do processo Y, visto que existe uma grande dificuldade para obter os resultados teóricos do processo Y, pois estes cálculos exigem o conhecimento de todas as taxas de atendimentos que estão na fila dos servidores e na ordem que chegaram.

Observação 5. *Observe que se a velocidade for maior do que zero, então o comprimento das filas tendem ao infinito. Entretanto, se as velocidade das filas são menores do que zero, isso significa que as filas zeram com probabilidade 1, isto é, o sistema é recorrente positivo (estável).*

Capítulo 4

Conclusões Finais

Este trabalho analisou o desempenho dos modelos de redes de filas com N servidores e $\mathcal{K}$ tipos de trabalho (ou de cliente) que apresentam as taxas dos atendimentos nos servidores que se alteram de acordo com os tipos de trabalho que eles podem processar, submetidos a três políticas de escolha de servidor: uma estática e duas dinâmicas, previamente estabelecidas.

As políticas estáticas são políticas de escolha de servidor que não consideram a configuração atual da rede no momento dos encaminhamentos dos trabalhos aos servidores. Enquanto que para as políticas dinâmicas a configuração atual da rede é fundamental no momento da escolha do servidor, ou seja, os clientes analisam a configuração atual da rede no momento em que chegam, e depois usando alguma política de escolha previamente estabelecida, encaminham-se para o servidor escolhido que processará o trabalho.

A análise com respeito as políticas estáticas foi estabelecida de acordo com a melhor política estática ($\mathrm{JFQ}_{\mathfrak{b}}$). Segundo a política $\mathrm{JFQ}_{\mathfrak{b}}$, os trabalhos são encaminhados com as probabilidades adequadas aos servidores, garantindo a estabilidade da rede.

As duas políticas dinâmicas analisadas neste trabalho foram: a política que encaminha os trabalhos aos servidores que apresentam a menor carga horária média total da trabalho na fila, ponderada pelo tempo médio do processamento do servidor escolhido (JSWA) e a política que encaminha os trabalhos aos servidores apresentam a menor carga horária média total de permanência do sistema, ponderada pelo tempo médio do processamento do servidor escolhido (JSWAS).

No entanto, por causa da grande dificuldade na análise do processo Y, pois este processo exige o conhecimento de todas as taxas dos atendimentos que estão na fila dos servidores e na ordem que chegaram, foi construído o processo X.

O processo X tem como objetivo analisar a rede de acordo com a carga horária média total de trabalho na fila dos servidores, sem a necessidade do conhecimento de todas as taxas dos atendimentos que estão na fila dos servidores e na ordem que chegaram. Além disso, através das simulações de dois exemplos de redes de filas apresentados no Capítulo 3, é possível perceber que o processo X apresenta um comportamento semelhante ao do processo Y.

Por isso, devido à simplicidade do processo X e por ser uma boa alternativa para analisar o processo Y, este trabalho analisou o processo Y utilizando os resultados obtidos pelo processo X. A fim de encontrar qual ou quais as políticas de escolha de servidor que oferecem um desempenho eficiente para as redes cujas taxas dos atendimentos dos servidores alteram-se de acordo com os tipos de trabalho que eles podem processar.

Entretanto, antes de tentar encontrar a política ideal de escolha de servidor baseada nas políticas previamente estabelecidas é necessário verificar se existe uma possível hierarquização dos servidores que constitui a rede.

Esta hierarquização consiste em agrupar todos os servidores que apresentam a mesma velocidade da fila num único agrupamento, de acordo com a ordem decrescente das velocidades dos servidores (ou filas) como estabelece o Teorema 2.2.1. E em seguida, procura-se a política ideal de escolha de servidor.

A política ideal de escolha de servidor é aquela política que oferece

- a velocidade da fila mais lenta é a mais rápida possível,
- o número de filas que crescem com velocidade máxima é o menor possível, e
- as velocidades das filas são decrescentes.

Com base nisso, é possível concluir que a política ideal de escolha de servidor para este modelo de redes de filas é a melhor política estática $\mathrm{JFQ_b}$.

Segundo as políticas dinâmicas analisadas, a política de escolha de servidor que considera a menor carga horária média total da fila ponderada pelo tempo médio do processamento do servidor escolhido (JSWA) apresentou a velocidade média

estimada da fila igual a velocidade média estimada que considera a menor carga horária média total do sistema ponderada pelo tempo médio do processamento do servidor escolhido (JSWAS).

No entanto, estas políticas dinâmicas não oferecem um desempenho tão eficiente quanto ao desempenho obtido pela política $\mathrm{JFQ}_{\mathtt{b}}$.

Portanto, é possível concluir que a melhor política estática $\mathrm{JFQ}_{\mathtt{b}}$ é uma política ideal de escolha de servidor para este modelo de redes de filas.

Referências Bibliográficas

[1] Andjel, E.D.; Menshikov, M.V.; Sisko. V. (2006) Positive recurrence of processes associated to crystal growth models. *Annals of Applied Probability*, Vol. 16, No. 3, 1059–1085.

[2] Andradóttir, S.; Ayhan, H. (2003) Dynamic server allocation for queueing networks with flexible servers. *Operations Research.*, Vol. 51, No. 6, 952–968.

[3] Andrade, E. (1998) Introdução à pesquisa operacional: métodos e modelos para a análise de decisão. 2ª edição, Rio de Janeiro, LTC.

[4] Asmussen, S. (2003) *Applied Probability and Queues*, *Springer-Verlag, NY.*

[5] Bramson, M. (1998) State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, Vol. 30, 89–148.

[6] Bramson, M. (2011) Stability of join the shortest queue networks. *Annals of Applied Probability*, Vol. 21, No. 4, 1568–1625.

[7] Dai, J.G.; Hasenbain, J.J.; Kim, B. (2007) Stability of Join-the-Shortest-Queue networks. *Queueing Systems*, Vol 57, 129–145.

[8] Fayolle, G.; Malyshev, V.A.; Menshikov, M.V. (1995) *Topics in Constructive Theory of Countable Markov Chains.* Cambridge University Press.

[9] Ford, L.R.; Fulkerson. (1974) *Flows in Networks.* Princeton University Pess,New Jersey.

[10] Foss, S.; Chernova, N. (1998) On the stability of partially accessible multi-station queue with state-dependent routing. *Queueing Systems*, Vol. 29, 55–73.

[11] Foley, R.D.; McDonald, D.R. (2001) Join the shortest queue: stability and exact asymptotics. *Annals of Applied Probability*, Vol. 11, 569–597.

[12] Kang, W.N.; Kelly, F.P.; Lee, N.H.;Williams, R.J. (2009) State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *The Annals of Probability*, Vol. 19, No. 5, 1719–1780.

[13] Luczak, M.J.; McDiarmid, C. (2006) On the maximum queue length in the supermarket model. *The Annals of Probability*, Vol. 34, No. 2, 493-527

[14] Menshikov, M.; Sisko, V.; Vachkovskaia, M. (2011) Introduction to Shape Stability for a Storage Model. *Methodology and Computing in Applied Probability,* Vol. 1, 1–22.

[15] Menshikov, M.; MacPhee, I.; Vachkovskaia, M. (2012) Dynamics of the non-homogeneous supermarket model. *Stochastic Models*, Vol. 28, 533–556.

[16] Menshikov, M.; Vachkovskaia, M.; Wade, A.R. (2008) Asymptotic behaviour of randomly reflecting billiards in unbounded tubular domains. *J. Stat. Phys.*, Vol. 132, 1097–1133.

[17] Mitzenmacher, M.; Richa, A.W.; Sitaraman, R. (2001) The power of two random choices: a survey of techniques and results, in Handbook of randomized computing, Vol. I, II, 255–312, *Comb. Optim.* **9**, Kluwer Acad. Publ., Dordrecht.

[18] Rockafellar, R.T. (1984) *Network Flows And Monotropic Optimization*, Pure and applied mathematics(New York),John Wiley & Sons,Inc, A Wiley-Interscience Publication.

[19] Ross, S.M. (2007) *Introduction to probability models*, 7^{a} ed. Academic Press (USA).

[20] Schilling, R.L. (2005) *Measures, integrals and martingales.* Cambridge University Press.

[21] Stolyar, A.L.; Tezcan, T.(2011) Shadow-routing based control of multi-server in overload. *Operations Research*, Vol. 59, No. 6, 1427–1444.

[22] Stolyar, A.L.; Tezcan, T. (2010) Control of systems with flexible multi-server pools: a shadow routing approach. *Queueing Systems.* Vol. 66, 1–51.

[23] Stolyar, A.L.; Yudovina, E. (2011) Instability of Natural Load Balancing in Large-Scale Flexible-Server Systems. Proceedings of the Forty-Ninth Allerton Conference, 361-368.

[24] Weber, R.R. (1978) On the optimal assignment of customers to parallel servers. *J. Appl. Probability*, Vol. 15, No. 2, 406–413.

[25] Vvedenskaya, N.D.; Dobrushin, R.L. F.I. Karpelevich (1996) A queueing system with a choice of the shorter of two queues—an asymptotic approach.*Problems Inform. Transmission*, Vol. 32, No. 1, 15–27.