Capítulo 4

Teoria de Filas

A Teoria de filas é uma das abordagens mais utilizadas no estudo de desempenho e dimensionamento de sistemas de comunicação de dados. Muita atenção deve ser dada aos processos de chegada e atendimento.

4.1 Definições básicas

Considere a Figura 4.1. Elementos chegam a uma fila com uma taxa (ou ritmo) de chegadas dada por λ. Os elementos são atendidos por M servidores com uma taxa de atendimento dada por μ.

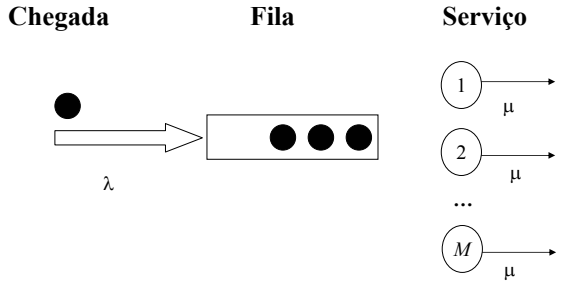


Figura 4.1: Seqüência de chamadas do sistema RPC

Podem-se definir as seguintes variáveis:

TA Tempo médio de atendimento, dado por $TA = 1/\mu$;

IC Intervalo médio entre chegadas, dado por $IC = 1/\lambda$;

TF Tempo médio gasto na fila;

TS Tempo médio gasto no sistema, dado por $TS = TF + TA$;

NA Número médio de elementos sendo atendidos;

NF Número médio de elementos na fila;

NS Número médio de elementos no sistema, dado por $NS = NF + NA$;

4.1.1 Leis de Little

As leis de Little aplicam-se a todos os sistemas de fila, independentemente do processo de chegada e atendimento:

$$NF = TF.\lambda \tag{4.1}$$

$$NS = TS.\lambda \tag{4.2}$$

4.1.2 Ocupação do Sistema

Define-se como ocupação (ou utilização) do sistema:

$$\rho = \frac{\lambda}{M.\mu} \tag{4.3}$$

Sendo que para que o sistema seja estacionário $\rho < 1$. Esta relação é fundamental para o estudo de sistemas de filas. Caso $\rho \geq 1$ a fila aumenta indefinidamente.

A ocupação do sistema não possui unidade. Significa a parcela do tempo em que os servidores estão atendendo requisições. Logo, o tempo livre dos servidores pode ser dado por $1 - \rho$.

> *Exercício 5:* Considere um sistema onde chegam requisições a um servidor. Abaixo estão os valores para os intervalos entre chegadas e tempo de atendimento para cada requisição:
>
> Intervalo entre Chegadas: 2,03; 3,50; 2,94; 2,26; 2,00; 2,22; 2,53; 2,69; 2,87; 3,33; 2,57; 2,46; 2,25; 2,14; 2,54.
>
> Tempo de Atendimento: 2,18; 1,70; 6,96; 3,86; 1,80; 1,75; 1,39; 1,41; 3,42; 1,63; 4,83; 3,51; 1,59; 1,83; 1,46.
>
> **Determine:**
>
> **a.** O ritmo médio de chegadas.
>
> **b.** O ritmo médio de atendimentos.

c. O tempo médio de espera na fila neste sistema (apenas observe a dinâmica da fila, não é necessário assumir nenhum modelo).

d. A ocupação do sistema.

e. O número médio de elementos na fila e no sistema.

□

Exercício 6: Suponha uma operadora de cartões de crédito. Muitos comerciantes ainda utilizam a modalidade de conexões discadas para acessar o sistema. Sabendo-se que o tempo médio de atendimento é de 20 segundos (tempo para enviar os dados necessários para processar a transação) e que o sistema tem uma taxa de chegada de 1000 conexões por minuto, determine a quantidade *mínima* de modems. □

4.2 Processos de Chegada e Atendimento

A identificação dos processos de chegada e atendimento permite a utilização de resultados bem estabelecidos para determinação dos valores de TF (e conseqüentemente de TS, NF e NS) para sistemas de filas. A identificação de tais processos utiliza os clássicos testes de aderência, estudados em probabilidade.

4.3 Notação de Kendall

É uma notação padrão para classificar sistemas de filas de acordo com as diferentes configurações possíveis.

A/B/C/K/P/Z

- `A` Distribuição do intervalo entre chegadas
- `B` Distribuição do tempo de serviço
- `C` Número de servidores
- `K` Número máximo de clientes no sistema (valor default ∞)
- `P` Tamanho da população (valor default ∞)
- `Z` Disciplina da fila (valor default FIFO)

As variáveis A e B podem assumir os seguintes valores:

M Distribuição exponencial (Markoviano);

D Determinístico;

E_k Distribuição de Erlang (k = shape parameter);

G Geral (qualquer distribuição)

Exemplo 7: $D/M/n$ descreve uma fila com o intervalo entre chegadas determinístico (sempre o mesmo intervalo), tempo de atendimento exponencial e n servidores. □

4.4 Resultados Clássicos

4.4.1 Modelo M/M/1

$$NF = \frac{\lambda^2}{\mu(\mu-\lambda)} \quad NS = \frac{\lambda}{\mu-\lambda}$$
$$TF = \frac{\lambda}{\mu(\mu-\lambda)} \quad TF = \frac{1}{\mu-\lambda}$$

Probabilidade de existirem n clientes no sistema: $P_n = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n$

4.4.2 Modelo M/M/c

$$TF = TA\frac{1-A}{M(1-\rho)(1-\rho A)}, \quad A = \frac{\sum_{i=0}^{M-1}\frac{(M\rho)^i}{i!}}{\sum_{i=0}^{M}\frac{(M\rho)^i}{i!}}$$

4.4.3 Modelo M/M/c com Perda de Chamada

Neste modelo o tamanho da fila é igual a zero. Caso todos os servidores estejam ocupados, as requisições são perdidas. A probabilidade de obter-se o sistema ocupado neste caso é dada por

$$\Pr\left[Ocupado\right] = \frac{\frac{(M\rho)^M}{M!}}{\sum_{i=0}^{M}\frac{(M\rho)^i}{i!}}$$

Em anexo pode ser encontrado um gráfico com a probabilidade de perda em função de ρ. Demais modelos podem ser encontrados em [Jain 1991] e [**?**].

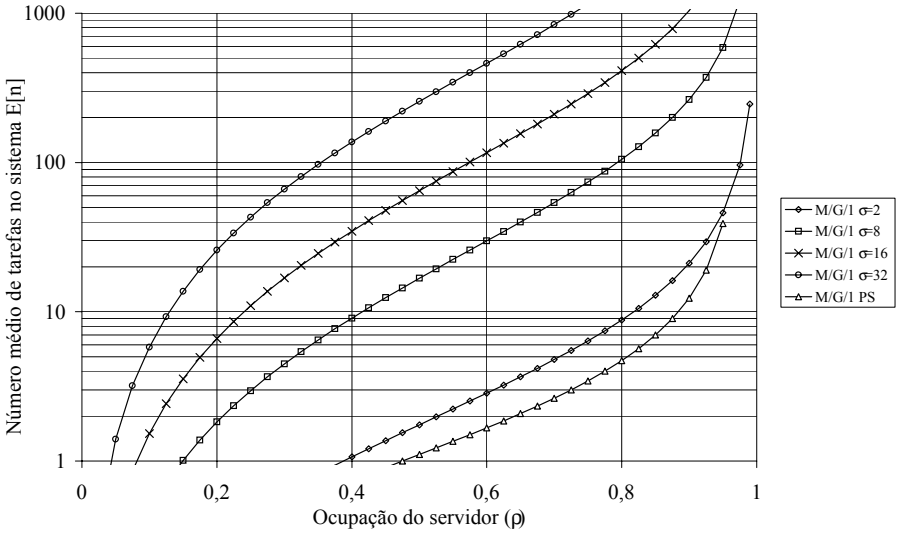


Figura 4.2: Número de tarefas no sistema com desvio padrão de $\sigma = 0, 2, 8, 16, 32$ no sistema *M/G/1* e no sistema *M/G/1 PS*

4.4.4 Modelo M/G/1

A equação que fornece o número médio de tarefas no sistema M/G/1, conhecida por equação de Pollaczek-Khinchin, é dada pela Equação 4.4,

$$E[n] = \frac{\rho^2}{2.(1-\rho)}.\left[1 + \frac{\sigma^2}{TA^2}\right] + \rho, \quad \rho < 1 \tag{4.4}$$

onde ρ é a ocupação do sistema, dado pela razão entre a taxa de chegada λ e a taxa de atendimento μ, $\rho = \lambda/\mu$. O desvio padrão do tempo de serviço é representado por σ e TA indica o tempo médio de serviço. O tempo total de atendimento pode ser calculado utilizando-se as leis operacionais de Little [Jain 1991].

Segundo [Jain 1991], a expressão que calcula o número de elementos no sistema *M/G/1 PS* é dada por $E[n] = \rho/(1-\rho)$. O tempo médio de resposta pode ser obtido com as leis de Little.

4.5 Exercícios Selecionados

Exercício 7: Considere um sistema onde chegam requisições a um servidor. Abaixo estão os valores para os intervalos entre chegadas e tempo de atendimento para cada requisição:

Intervalo entre Chegadas: 2,0-1,5-2,9-2,2-3,0-4,2-2,5-5,6-2,8-3,3-2,5-2,4-2,2-2,1-2,5 Tempo de Atendimento: 2,1-1,7-7,9-3,8-1,8-1,7-1,3-1,4-2,4-1,6-3,8-1,5-1,5-1,1-0,4

1. Qual o ritmo médio de chegadas?
2. Qual o ritmo médio de atendimentos?
3. Qual o tempo médio de espera na fila neste sistema? (apenas observe a dinâmica da fila, não é necessário assumir nenhum modelo)

□

Exercício 8: Foi observado o comportamento de um certo servidor de banco de dados durante um período de tempo onde o sistema era estacionário.

Durante o período de 1 minuto de observação, o tempo livre do sistema (idle time) foi de 10 segundos. A taxa média de chegada foi de 5 requisições por segundo.

Utilizando o modelo M/M/1, determine:

1. A utilização do sistema;
2. O número de médio requisições sendo processadas ;
3. O tempo médio de resposta;
4. A probabilidade do número de tarefas no sistema ser maior que 10.

Já é conhecido o fato de que os tempos de resposta em servidores seguem distribuições de cauda pesada. Recalcule os ítens acima considerando o sistema como uma fila M/G/1, aplicando valores crescentes para o desvio padrão do tempo de serviço. O que ocorre com o tempo médio de resposta a medida que a variação do tempo de serviço aumenta? □

Exercício 9: Um sistema com uma base de dados consiste de 2 discos rígidos compartilhando uma fila única. Suponha que cada os servidores estão configurados em RAID-1 (espelhamento de informações nos discos), ou seja, ambos os discos contem a mesma informação. O tempo médio para execução de uma operação de E/S pelo disco é de 50m segundos. As requisições de E/S chegam ao sistema a uma taxa de 30 requisições por segundo. Utilizando o modelo M/M/3, determine o seguinte:

1. A utilização média dos discos rígidos;
2. A probabilidade do sistema esta ocioso;

3. O número médio de requisições de acesso no sistema e o número médio de requisições esperando na fila;
4. O tempo médio de resposta.

□

Exercício 10: Suponha que no problema anterior os discos contenham informações diferentes. Recalcule qual seria oa tempo de resposta nesta situação. □

Exercício 11: Um banco possui dois funcionários trabalhando no setor de atendimento ao público. O primeiro trabalha apenas com depósitos e o segundo, com retiradas. Sabe-se que ambos atendem uma média de 3 minutos por cliente (a distribuição é desconhecida), com um desvio padrão de 5 minutos. Suponha que a chegada e o atendimento são processos markoviandos. com média de 16 chegadas por hora para os depositantes e 14 chegadas por hora para os que vão fazer retirada. Qual seria o efeito no tempo médio no sistema se ambos os funcionários trabalhassem tanto com retiradas como com depósitos? □

Exercício 12: Suponha que um sistema computacional foi estruturado da seguinte maneira: um cluster composto de 3 servidores recebe uma requisição, processa e encaminha para um dos dois servidores de banco de dados que contem a informação desejada. O cluster foi criado porque a capacidade de processamento dos computadores que processam a requisição é pequena em relação aos servidores de banco de dados. Suponha que o tempo médio de atendimento de uma requisição por um dos computadores do cluster é de 2 segundos. O tempo médio de atendimento de uma requisição para o servidor de banco de dados é de 0,5 segundo. O sistema recebe 3600 requisições por hora. Considere que os processos de chegada e atendimento são Markovianos. Utilizando a teoria de filas, determine:

1. Ocupação de cada servidor;
2. Tamanho médio da fila em cada um dos servidores e total;
3. Tempo médio de resposta em cada um dos servidores e total.

□

Exercício 13: Um provedor de acesso à internet possui 52.291 clientes. A taxa média de chegada é de 40 ligações por hora. Sabendo-se que a tempo médio de conexão é de 10 minutos, e que ambos seguem a distribuição exponencial, qual será a quantidade de modems necessários para que a probabilidade de perda de ligação seja menor que 3%? □

Exercício 14: Considere um servidor Web, com os seguintes parâmetros de desempenho:

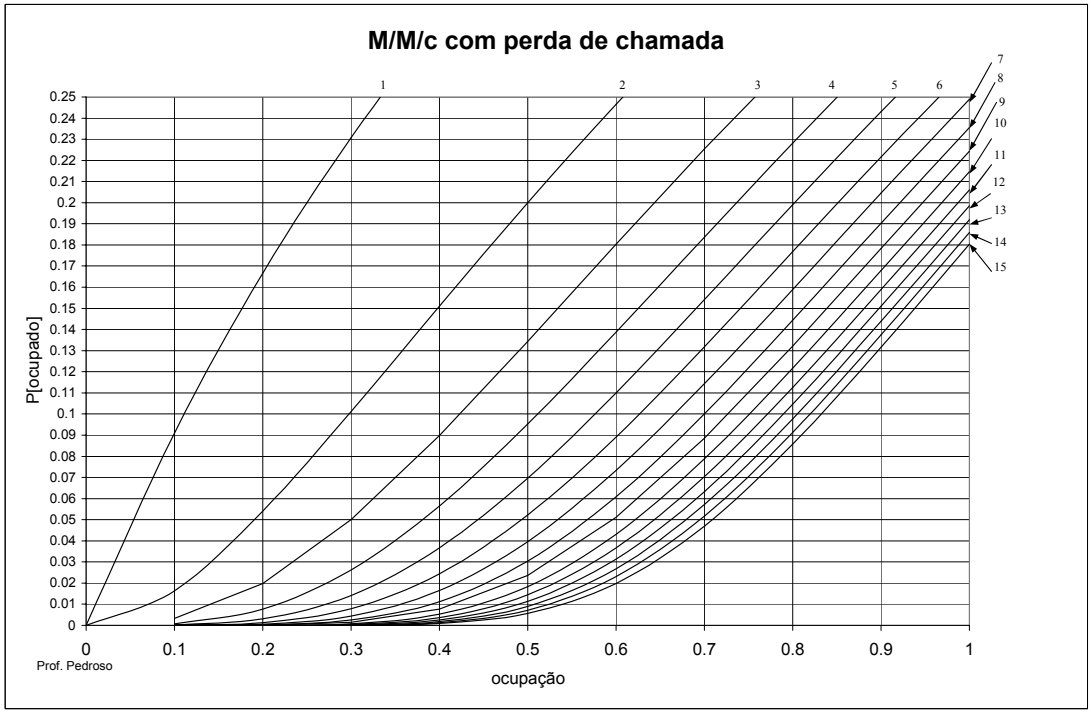
- O tempo médio que um servidor leva para tratar uma requisição é de 100ms.
- Na HMM, a taxa média de chegada de requisições é de 1100 requisições por minuto.
- Deseja-se utilizar um cluster de servidores para o serviço.

Considerando que as chegadas e atendimentos são processos Markovianos, determine:

1. Qual o número mínimo de servidores?
2. Qual o número de servidores para que a ocupação média do sistema seja de 50%?

□

Exercício 15: No problema anterior, foi observado que o tempo de processamento da requisição segue uma distribuição de cauda pesada. Explique qual será a consequêcia na previsão realizada no exercício anterior. □


M/M/c com perda de chamada
P[ocupado]
ocupação
Prof. Pedroso
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1