

## Algoritmos para Projeto de Redes de Telecomunicações: Fila M/M/m

Programa de Pós-Graduação em Informática (PPGIA) - PUCPR

Prof. Marcelo E. Pellenz  
<http://www.ppgia.pucpr.br/~marcelo>  
[marcelo@ppgia.pucpr.br](mailto:marcelo@ppgia.pucpr.br)

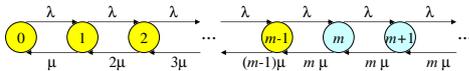
## Fila M/M/m

- Aplicações
  - Modelagem de sistemas ou dispositivos multiprocessados que possuem vários servidores idênticos.
- Considerações
  - Todas as tarefas (clientes) esperando pelo atendimento dos servidores estão em uma única fila.
  - Assumir  $m$  servidores, cada um com taxa de serviço de  $\mu$

2

## Fila M/M/m

- Existem  $m$  servidores idênticos e uma única fila:



- A fila é modelada como um processo de nascimento/morte onde:

$$\lambda_n = \lambda, \quad n = 0, 1, 2, \dots, \infty$$

$$\mu_n = \begin{cases} n\mu, & n = 1, 2, \dots, m-1 \\ m\mu, & n = m, m+1, \dots, \infty \end{cases}$$

3

## Teorema

- A probabilidade em regime permanente,  $p_n$ , de que um processo de nascimento e morte esteja no estado  $n$  é dada por:

$$p_n = \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n} p_0, \quad n = 1, 2, \dots, \infty$$

- A probabilidade  $p_0$  é a probabilidade de que o sistema se encontre no estado zero (vazio).

4

## Fila M/M/m

- Utilizando o teorema, obtemos:

$$p_n = \begin{cases} \frac{\lambda^n}{n! \mu^n} p_0, & n = 1, 2, \dots, m-1 \\ \frac{\lambda^n}{m! m^{n-m} \mu^n} p_0, & n = m, m+1, \dots, \infty \end{cases}$$

- Em termos da utilização de cada servidor  $\rho = \lambda/m\mu$ :

$$p_n = \begin{cases} \frac{(m\rho)^n}{n!} p_0, & n = 1, 2, \dots, m-1 \\ \frac{\rho^n m^m}{m!} p_0, & n = m, m+1, \dots, \infty \end{cases}$$

5

## Fila M/M/m

- A probabilidade de não haver clientes no sistema é calculada pela relação:

$$\sum_{n=0}^{\infty} p_n = 1$$

- Esta relação resulta em:

$$p_0 + p_0 \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} + p_0 \frac{(m\rho)^m}{m!} \sum_{n=m}^{\infty} \rho^{n-m} = 1$$

$$p_0 = \left[ 1 + \frac{(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

6

## Fila M/M/m

- A probabilidade de que um cliente que chega tenha que esperar na fila (todos os servidores ocupados) é dada por:

$$\begin{aligned} C(m, \rho) &= P(\geq m \text{ jobs}) = p_m + p_{m+1} + p_{m+2} + \dots \\ &= p_0 \frac{(m\rho)^m}{m!} \sum_{n=m}^{\infty} \rho^{n-m} \\ &= p_0 \frac{(m\rho)^m}{m!(1-\rho)} \end{aligned}$$

- Para  $m=1$  (Servidor Único - Fila M/M/1),  $C(m, \rho)$  é dada por  $\rho$ .
- Esta distribuição é conhecida como **Fórmula Erlang-C**

7

## Distribuição Erlang-C

- Aplicação no dimensionamento de *call center*
  - $C(c, a)$  = Probabilidade de Atrasar uma Chamada
  - $c$  = Número de Circuitos
  - $a$  = Carga Total de Tráfego

$$C(c, a) = \frac{a^c \cdot c}{c!(c-a)} \quad p_0 = \left[ 1 + \frac{(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

$$C(m, \rho) = p_0 \frac{(m\rho)^m}{m!(1-\rho)}$$

8

## Distribuição Erlang-C

$$U = m\rho = \frac{\lambda}{\mu} \quad p_0 = \left[ \frac{U^m}{m!(1-\frac{U}{m})} + \sum_{n=0}^{m-1} \frac{U^n}{n!} \right]^{-1}$$

$$C(m, U) = p_0 \frac{U^m}{m!(1-\frac{U}{m})} = \frac{U^m}{\frac{U^m}{m!(1-\frac{U}{m})} + \sum_{n=0}^{m-1} \frac{U^n}{n!}} = \frac{\frac{m \cdot U^m}{m!(m-U)}}{\frac{U^m}{m!(1-\frac{U}{m})} + \sum_{n=0}^{m-1} \frac{U^n}{n!}}$$

9

## Distribuição Erlang-C

$$C(c, a) = \frac{a^c \cdot c}{c!(c-a)} = \frac{U^m}{\frac{U^m}{m!(1-\frac{U}{m})} + \sum_{n=0}^{m-1} \frac{U^n}{n!}}$$

$$C(m, U) = p_0 \frac{U^m}{m!(1-\frac{U}{m})} = \frac{\frac{m \cdot U^m}{m!(m-U)}}{\frac{U^m}{m!(1-\frac{U}{m})} + \sum_{n=0}^{m-1} \frac{U^n}{n!}}$$

10

## Número Médio de Clientes em Espera

- Número Médio de Clientes em Espera

$$\begin{aligned} E[n_q] &= \sum_{n=m+1}^{\infty} (n-m)p_n = p_0 \frac{(m\rho)^m}{m!} \sum_{n=m+1}^{\infty} (n-m)\rho^{n-m} \\ &= p_0 \frac{(m\rho)^m \rho}{m!(1-\rho)^2} = \frac{\rho \cdot C(m, \rho)}{1-\rho} \\ E[n_q] &= \frac{\rho \cdot C(m, \rho)}{1-\rho} \end{aligned}$$

11

## Número Médio de Clientes em Atendimento

$$\begin{aligned} E[n_s] &= \sum_{n=1}^{m-1} n \cdot p_n + \sum_{n=m}^{\infty} m \cdot p_n \\ &= 1p_0 \frac{(m\rho)}{1!} + 2p_0 \frac{(m\rho)^2}{2!} + \dots + (m-1)p_0 \frac{(m\rho)^{m-1}}{(m-1)!} \\ &\quad + m(p_m + p_{m+1} + p_{m+2} + \dots) \\ &= m\rho \left( p_0 + p_0 \frac{(m\rho)}{1!} + p_0 \frac{(m\rho)^2}{2!} + \dots + p_0 \frac{(m\rho)^{m-2}}{(m-2)!} \right) + mC(m, \rho) \\ &= m\rho(p_0 + p_1 + p_2 + \dots + p_{m-2}) + mC(m, \rho) \\ &= m\rho[1 - p_{m-1} - C(m, \rho)] + mC(m, \rho) \\ &= m\rho - m\rho p_{m-1} + mC(m, \rho)(1-\rho) \\ &= m\rho \quad \text{dado que } mC(m, \rho)(1-\rho) = m\rho p_{m-1} \end{aligned}$$

12

## Número de Clientes no Sistema

- Número médio de clientes em Atendimento:  $E[n_s] = m \rho$
- Número médio de clientes no sistema:

$$E[n] = E[n_q] + E[n_s] = \frac{\rho \cdot C(m, \rho)}{1 - \rho} + m\rho$$

- Pode-se mostrar que a variância de  $n$  e  $n_q$  são dadas por:

$$\text{Var}[n] = m\rho + \rho C(m, \rho) \left[ \frac{1 + \rho - \rho C(m, \rho)}{(1 - \rho)^2} + m \right]$$

$$\text{Var}[n_q] = \frac{C(m, \rho) \rho [1 + \rho - \rho C(m, \rho)]}{(1 - \rho)^2}$$

13

## Tempos Médios

- Tempo médio de resposta usando a Lei de Little:

$$E[r] = \frac{E[n]}{\lambda} = \frac{1}{\mu} + \frac{C(m, \rho) / m\mu}{1 - \rho} = \frac{1}{\mu} \left( 1 + \frac{C(m, \rho)}{m(1 - \rho)} \right)$$

- Do mesmo modo, o tempo médio de espera em fila é dado por:

$$E[w] = \frac{E[n_q]}{\lambda} = \frac{C(m, \rho)}{m\mu(1 - \rho)}$$

14

## Tempo de Resposta

- Função Distribuição de Probabilidade do tempo de resposta:

$$F[r] = \begin{cases} 1 - e^{-\mu r} - \frac{C(m, \rho)}{1 - m + m\rho} e^{-m\mu(1-\rho)r} - e^{-\mu r}, & \rho \neq (m-1)/m \\ 1 - e^{-\mu r} - C(m, \rho)\mu r e^{-\mu r}, & \rho = (m-1)/m \end{cases}$$

- O tempo de resposta,  $r$ , não é exponencialmente distribuído, a não ser que  $m=1$ .
- Em geral, o coeficiente de variação, isto é, a razão entre o desvio padrão e a média de  $r$  é menor do que 1.

15

## Tempo de Espera

- A função distribuição de probabilidade do tempo de espera é dado por:

$$F(w) = 1 - C(m, \rho) e^{-m\mu(1-\rho)w}$$

- Dado que  $w$  possui uma função distribuição exponencial truncada, o posto percentil  $q$  pode ser calculado do seguinte modo:

$$w_q = \max \left\{ 0, \frac{1}{m\mu(1-\rho)} \cdot \ln \left( \frac{100 \cdot C(m, \rho)}{100 - q} \right) \right\}$$

16

## Exercício 1

- Estudantes chegam a um laboratório de computação de acordo com a distribuição de Poisson a uma taxa média de 10 estudantes/hora. Cada estudante gasta em média 20 minutos no computador e assume-se que este tempo seja exponencialmente distribuído. O laboratório tem atualmente 5 computadores e alguns alunos têm reclamado que os tempos de espera são muito longos.
- a) Analise o laboratório usando um modelo de filas MM/5
  - Taxa de Chegada \ Taxa de Atendimento \ Utilização Efetiva
  - Probabilidade de todos os terminais estarem desocupados
  - Probabilidade de todos os terminais estarem ocupados
  - Número médio de estudantes no laboratório (sistema)
  - Número médio de estudantes esperando na fila
  - Número médio de estudantes usando os computadores (Atendimento)
  - Média e Variância do tempo gasto no laboratório
  - Tempo médio de espera
  - O percentil 90 do tempo de espera

17

## Exercício 1

- b) Os estudantes querem limitar o seu tempo de espera para uma média de 2 minutos e não mais do que 5 minutos em 90% dos casos.
- c) Isso é viável? Se for, quantos computadores seriam necessários?

18

### Exercício 1

- a) O laboratório pode ser modelado como um sistema de fila M/M/5:
- Taxa de Chegada:

$$\lambda \approx \frac{10}{60} = 0,167 \text{ estudantes/minuto}$$

- Taxa de Atendimento:

$$\mu = \frac{1}{20} = 0,05 \text{ estudantes/minuto}$$

- Intensidade de Tráfego:

$$\rho = \frac{\lambda}{m\mu} = \frac{0,167}{5 \cdot 0,05} = 0,67$$

19

### Exercício 1

- a) O laboratório pode ser modelado como um sistema de fila M/M/5:
- Probabilidade de todos os terminais estarem *desocupados*:

$$p_0 = \left[ 1 + \frac{(5 \cdot 0,67)^1}{5!(1-0,67)} + \frac{(5 \cdot 0,67)^2}{2!} + \frac{(5 \cdot 0,67)^3}{3!} + \frac{(5 \cdot 0,67)^4}{4!} \right]^{-1} = 0,0318$$

- Probabilidade de todos os terminais estarem *ocupados*:

$$C(m, \rho) = p_0 \frac{(m\rho)^m}{m!(1-\rho)} = 0,0318 \frac{(5 \cdot 0,67)^5}{5!(1-0,67)} = 0,33$$

- Utilização média de cada terminal:

$$\rho = 0,67$$

20

### Exercício 1

- a) O laboratório pode ser modelado como um sistema de fila M/M/5:
- Número médio de estudantes no laboratório (sistema):

$$E[n] = m\rho + \frac{\rho \cdot C(m, \rho)}{1-\rho} = 5 \cdot 0,67 + \frac{0,67 \cdot 0,33}{1-0,67} = 4$$

- Número médio de estudantes esperando na fila:

$$E[n_q] = \frac{\rho \cdot C(m, \rho)}{1-\rho} = \frac{0,67 \cdot 0,33}{1-0,67} = 0,65$$

- Número médio de estudantes usando os computadores (Atendimento):

$$E[n_s] = E[n] - E[n_q] = 4 - 0,65 = 3,35$$

21

### Exercício 1

- a) O laboratório pode ser modelado como um sistema de fila M/M/5:
- Média e Variância do tempo gasto no laboratório:

$$E[r] = \frac{1}{\mu} \left( 1 + \frac{C(m, \rho)}{m(1-\rho)} \right) = \frac{1}{0,05} \left( 1 + \frac{0,33}{5 \cdot (1-0,67)} \right) = 24$$

$$\text{Var}[r] = \frac{1}{\mu^2} \left( 1 + \frac{C(m, \rho)(2 - C(m, \rho))}{m^2(1-\rho)^2} \right) = \frac{1}{0,05^2} \left( 1 + \frac{0,33 \cdot (2-0,33)}{5^2 \cdot (1-0,67)^2} \right) = 479$$

- Tempo médio de espera:

$$E[w] = \frac{C(m, \rho)}{m\mu(1-\rho)} = \frac{0,33}{5 \cdot 0,05 \cdot (1-0,67)} = 4$$

22

### Exercício 1

- a) O laboratório pode ser modelado como um sistema de fila M/M/5:
- O percentil 90 do tempo de espera é:

$$w_q = \max \left\{ 0, \frac{1}{m\mu(1-\rho)} \cdot \ln \left( \frac{100 \cdot C(m, \rho)}{100 - q} \right) \right\} = 14$$

- Isso significa que **10%** dos estudantes tem que esperar **mais que 14 minutos**.

23

### Exercício 1

- b) Os estudantes querem limitar o seu tempo de espera para uma média de 2 minutos e não mais do que 5 minutos em 90% dos casos.
- Analisando o sistema com  $m=6$  computadores, mantendo as mesmas taxas de chegada e atendimento de  $\lambda=0,167$  e  $\mu=0,05$ , temos:
- Intensidade de Tráfego:  $\rho = 0,556$
- Probabilidade de todos os terminais estarem desocupados:  $p_0 = 0,0346$
- Probabilidade de todos os terminais estarem ocupados:  $C(m, \rho) = 0,15$
- Tempo médio de espera:  $E[w_q] = 1,1$  minutos
- O percentil 90 do tempo de espera:  $w_q = 3$

24

## M/M/1 vs. M/M/m

- Com  $m$  servidores, qual a melhor alternativa ?
  - Manter **filas separadas** para cada servidor
  - Manter **uma única fila** para todos os servidores
- Para chegadas de Poisson e tempos de serviço exponenciais:
  - $m$  filas M/M/1 com taxa de chegada  $\lambda/m$
  - Uma única fila M/M/m com taxa de chegada  $\lambda$
- Vamos verificar que uma única fila é melhor quando as tarefas são homogêneos.

25

## Exemplo

- No exemplo anterior, considere que os 5 computadores estão localizados em 5 diferentes unidades do campus, portanto é necessário manter filas separada para cada um.
- Neste caso, o sistema é modelado como 5 filas M/M/1 separadas. Usando  $m=1$ ,  $\lambda=0,167/5=0,0333$  e  $\mu=0,05$ , temos:

$$\text{Intensidade de tráfego } \rho = \frac{0,0333}{0,05} = 0,67$$

$$E[r] = \frac{1/\mu}{1-\rho} = \frac{1/0,05}{1-0,67} = 60$$

$$\text{Var}[r] = \frac{1/\mu^2}{(1-\rho)^2} = \frac{1/0,05^2}{(1-0,67)^2} = 3600$$

M/M/5  
24

479

26

## Exercício 2

- Num dado sistema, 24 terminais compartilham uma linha de transmissão de 9600 bps. Cada terminal envia 10 msgs/min. O tamanho das mensagens é exponencialmente distribuídas com um tamanho médio de 2000 bits.
- a) Qual o tempo médio que uma mensagem gasta no sistema ?
- b) Suponha que este tempo seja inaceitável. Uma solução possível é utilizar um canal adicional de 9600 bps e colocar 12 terminais em cada canal. Qual o tempo total no sistema agora ?
- c) Se os dois canais forem utilizados como um sistema de fila M/M/2, qual o tempo total no sistema ?
- d) Se os dois canais forem utilizados como um sistema de fila M/M/1, com taxa de 19200bps, qual o tempo total no sistema ?

27

## Fila M/M/∞

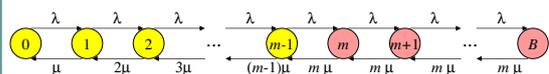
- Número infinito de servidores.
- É um caso especial das filas M/M/m.
- Os clientes nunca têm que esperar, pois há sempre um servidor disponível.
- O tempo de resposta é igual ao tempo de serviço.
- São também chamadas de **centros de atraso**.
- São utilizadas para representar recursos dedicados, tais como terminais num sistema de tempo compartilhado.
- Suas equações podem ser derivadas das utilizadas para filas M/M/m.

28

## Fila M/M/m/B

## Fila M/M/m/B (Buffer Finito)

- Número de posições na fila,  $B$  ( $\geq m$ ), é finito
- Quando todas as posições estiverem ocupadas, as novas chegadas serão perdidas:



- Modelada como um processo de nascimento e morte, onde:

$$\lambda_n = \lambda, \quad n = 0, 1, 2, \dots, B-1$$

$$\mu_n = \begin{cases} n\mu, & n = 1, 2, \dots, m-1 \\ m\mu, & n = m, m+1, \dots, B \end{cases}$$

30

## Fila M/M/m/B

- Utilizando o Teorema, obtém-se:

$$p_n = \begin{cases} \frac{\lambda^n}{n! \mu^n} p_0, & n = 1, 2, \dots, m-1 \\ \frac{\lambda^n}{m! m^{n-m} \mu^n} p_0, & n = m, m+1, \dots, B \end{cases}$$

- Em termos da intensidade de tráfego  $\rho = \lambda/m\mu$ :

$$p_n = \begin{cases} \frac{(m\rho)^n}{n!} p_0, & n = 1, 2, \dots, m-1 \\ \frac{\rho^n m^m}{m!} p_0, & n = m, m+1, \dots, B \end{cases}$$

31

## Fila M/M/m/B

- A probabilidade de **não haver clientes** no sistema é calculada pela relação:

$$\sum_{n=0}^B p_n = 1$$

- Esta relação em:

$$p_0 + p_0 \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} + p_0 \frac{(m\rho)^m}{m!} \sum_{n=m}^B \rho^{n-m} = 1$$

$$p_0 = \left[ 1 + \frac{(1 - \rho^{B-m+1})(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

32

## Número Médio de Clientes

$$E[n] = \sum_{n=1}^B n p_n$$

$$E[n_q] = \sum_{n=m+1}^B (n-m) p_n$$

- A variância e outras estatísticas de  $n$  e  $n_q$  podem ser calculadas de forma semelhante.

33

## Taxa Efetiva de Chegadas

- Todas as chegadas que ocorrem enquanto o sistema está cheio ( $n=B$ ) são perdidas.
- A taxa de tarefas que efetivamente entram no sistema é dada por:

$$\lambda' = \sum_{n=0}^{B-1} \lambda \cdot p_n = \lambda \sum_{n=0}^{B-1} p_n = \lambda(1 - p_B)$$

- A diferença  $\lambda - \lambda' = \lambda p_B$  representa a **taxa de perda de pacotes**.

34

## Tempos Médios

- Tempo médio de resposta usando a Lei de Little:

$$E[r] = \frac{E[n]}{\lambda'} = \frac{E[n]}{\lambda(1-p_B)}$$

- Do mesmo modo, o tempo médio de espera é dado por:

$$E[w] = \frac{E[n_q]}{\lambda'} = \frac{E[n_q]}{\lambda(1-p_B)}$$

35

## Utilização da Fila M/M/m/B

- Se observarmos o sistema por um tempo longo, por exemplo,  $T$  segundos:

- O número total de clientes chegando e recebendo serviço será  $\lambda'T$
- O tempo total ocupado dos  $m$  servidores para atender estes clientes será de  $\lambda'T\mu$
- A utilização de cada servidor será dada por:

$$U = \frac{\text{tempo ocupado por servidor}}{\text{tempo total}}$$

$$= \frac{(\lambda'T / \mu) / m}{T} = \frac{\lambda'}{m\mu} = \rho(1 - p_B)$$

- $p_B$  é a probabilidade de o sistema estar cheio

36

## Fórmula de Perdas de Erlang

- Em um sistema  $M/M/m/m$ , o tamanho do buffer é exatamente igual ao número de servidores. A probabilidade de perda é dada por:

$$p_m = \frac{(m\rho)^m}{m!} p_0 = \frac{(m\rho)^m / m!}{\sum_{j=0}^m [(m\rho)^j / j!]}$$

- As fórmulas aqui apresentadas podem ser reduzidas ao caso de  $m=1$  servidor, caracterizando um sistema  $M/M/1/B$ .