Análise de uma Fila Única

"The Art of Computer Systems
Performance Analysis"
Raj Jain, Cap. 31

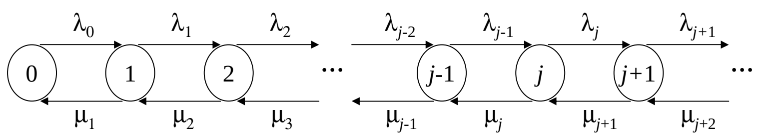
Fila Única

- O modelo de filas mais simples contém apenas uma fila
- Pode ser usado para analisar recursos individuais em sistemas de computação
- Muitas filas podem ser modeladas como processos de nascimento e morte

Processos de Nascimento e Morte

- Um processo de nascimento e morte é útil para modelar sistemas nos quais os jobs chegam um de cada vez (e não em lotes)
- O estado do sistema pode ser representado pelo número de jobs n no sistema
- A chegada de um novo job (nascimento) leva o estado para $n+1$
- A partida de um job (morte) leva o estado para $n-1$

Diagrama de Transição de Estados



- Quando o sistema está no estado n, significa que ele possui n jobs
- Taxa de novas chegadas: λ_n
- Taxa de serviço/atendimento: μ_n
- Assume-se que tanto os intervalos entre as chegadas quanto os tempos de serviço são exponencialmente distribuídos

Probabilidade de Estados

- Teorema 31.1: A probabilidade em regime permanente p_n de que um processo de nascimento e morte esteja no estado n é dada pelo teorema:

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

  - onde p_0 é a probabilidade de que o sistema se encontre no estado 0 (vazio)

Prova 31.1

- Suponha que o sistema se encontre no estado j no instante t
- Seja Δt o próximo intervalo de tempo, com duração muito curta, de modo que não haja dois eventos no mesmo intervalo
- No próximo intervalo de tempo, o sistema pode transitar para o estado $j - 1$ ou $j + 1$ com as seguintes probabilidades:

$$P\{n(t+\Delta t) = j+1 \mid n(t) = j\} = \text{probabilidade de uma chegada em } \Delta t$$
$$= \lambda_j \Delta t$$
$$P\{n(t+\Delta t) = j-1 \mid n(t) = j\} = \text{probabilidade de uma partida em } \Delta t$$
$$= \mu_j \Delta t$$
$$P\{n(t+\Delta t) = j \mid n(t) = j\} = \text{probabilidade de permanecer em } j$$
$$= 1 - \lambda_j \Delta t - \mu_j \Delta t$$

Prova 31.1

- Chamando de $p_j(t)$ a probabilidade de estar no estado j no instante t, podemos escrever o seguinte conjunto de equações lineares:

$$
\begin{aligned}
p_0(t+\Delta t) &= (1-\lambda_0\Delta t)\,p_0(t) + \mu_1\Delta t\ p_1(t)\\
p_1(t+\Delta t) &= \lambda_0\Delta t\ p_0(t) + (1-\mu_1\Delta t - \lambda_1\Delta t)\,p_1(t) + \mu_2\Delta t\ p_2(t)\\
p_2(t+\Delta t) &= \lambda_1\Delta t\ p_1(t) + (1-\mu_2\Delta t - \lambda_2\Delta t)\,p_2(t) + \mu_3\Delta t\ p_3(t)\\
&\ \vdots\\
p_j(t+\Delta t) &= \lambda_{j-1}\Delta t\ p_{j-1}(t) + (1-\mu_j\Delta t - \lambda_j\Delta t)\,p_j(t) + \mu_{j+1}\Delta t\ p_{j+1}(t)
\end{aligned}
$$

Prova 31.1

- A j-ésima equação pode ser escrita da seguinte forma:

$$
\lim_{\Delta t\to 0}\frac{p_j(t+\Delta t)-p_j(t)}{\Delta t} = \lambda_{j-1}p_{j-1}(t) - (\mu_j+\lambda_j)\,p_j(t) + \mu_{j+1}\ p_{j+1}(t)
$$

$$
\frac{d\ p_j(t)}{dt} = \lambda_{j-1}p_{j-1}(t) - (\mu_j+\lambda_j)\,p_j(t) + \mu_{j+1}\ p_{j+1}(t)
$$

- Em regime permanente, $p_j(t)$ aproxima-se do valor constante p_j, isto é:

$$
\lim_{\Delta t\to 0} p_j(t) = p_j
$$

e

$$
\lim_{\Delta t\to 0}\frac{d\ p_j(t)}{dt} = 0
$$

Prova 31.1

- Substituindo estes valores na *j*-ésima equação, obtemos:

$$0 = \lambda_{j-1} p_{j-1} - (\mu_j + \lambda_j) p_j + \mu_{j+1} \; p_{j+1}$$

$$p_{j+1} = \left(\frac{\mu_j + \lambda_j}{\mu_{j+1}} \right) p_j - \frac{\lambda_{j-1}}{\mu_{j+1}} p_{j-1}, \qquad j = 1,2,3,\ldots$$

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

- A solução para este conjunto de equações é a seguinte:

$$p_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} p_0 = p_0 \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}}, \qquad n = 1,2,\ldots,\infty \quad (31.1)$$

P_0

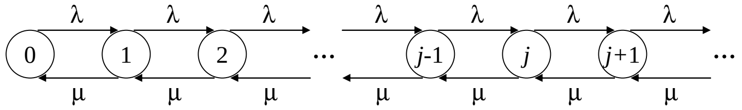
- O Teorema 31.1 permite determinar a probabilidade de equilíbrio p_n em termos de p_0. Usando a condição adicional de que a soma de todas as probabilidades deve ser 1, obtemos:

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \left[\lambda_j / \mu_{j+1} \right]}$$

- A partir dessas probabilidades de estado, podemos calcular muitas outras medidas de desempenho, para diferentes tipos de sistemas de filas. Ex: M/M/1, M/M/m, M/M/m/K

Fila M/M/1

Fila M/M/1



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

$$\lambda_n = \lambda, \qquad n = 0,1,2,\ldots,\infty$$

$$\mu_n = \mu, \qquad n = 1,2,\ldots,\infty$$

- Utilizando o Teorema 31.1, obtemos:

$$p_n = \underbrace{\left(\frac{\lambda}{\mu}\right)}_{\rho = \text{ intensidade de tráfego}}^{n} p_0, \quad n = 1,2,\ldots,\infty$$

Fila M/M/1

- Portanto:

$$p_n = \rho^n p_0, \quad n = 1,2,\ldots,\infty$$

- Como a soma das probabilidades deve ser igual 1, obtemos:

$$p_0 = \frac{1}{1+\rho+\rho^2+\cdots+\rho^\infty} = 1-\rho$$

- Esta soma infinita é a soma de uma série geométrica, que só converge se $\lambda/\mu < 1$ (equilíbrio!)

- Substituindo na expressão para p_n, obtemos:

$$p_n = (1-\rho)\rho^n, \quad n = 0,1,2,\ldots,\infty$$

Outras Propriedades da Fila M/M/1

- **Utilização do servidor** (probabilidade de ter um ou mais jobs no sistema):

$$U = 1 - p_0 = \rho$$

- **Número médio de jobs no sistema**:

$$E[n] = \sum_{n=1}^{\infty} n p_n = \sum_{n=1}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}$$

- **Variância do número de jobs no sistema**:

$$\mathrm{Var}[n] = E[n^2] - (E[n])^2 = \left(\sum_{n=1}^{\infty} n^2(1-\rho)\rho^n\right) - (E[n])^2 = \frac{\rho}{(1-\rho)^2}$$

Outras Propriedades da Fila M/M/1

- **Probabilidade de ter *n* ou mais jobs no sistema**:

$$P(\geq n \text{ jobs no sistema}) = \sum_{j=n}^{\infty} p_j = \sum_{j=n}^{\infty} (1-\rho)\rho^j = \rho^n$$

- O **Tempo médio de resposta** pode ser calculado usando a Lei de Little:

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

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

Outras Propriedades da Fila M/M/1

- A **função de distribuição de probabilidade acumulada (CDF),** $F_x(a) = P(x \leq a)$, do tempo de resposta é dada por:

$$F(r) = 1 - e^{-r\mu(1-\rho)}$$

- Note que o tempo de resposta é distribuído exponencialmente

- **Posto percentil *q*** :

$$1 - e^{-r_q \mu(1-\rho)} = \frac{q}{100}$$

$$r_q = \frac{1}{\mu(1-\rho)} \ln\left(\frac{100}{100-q}\right)$$

Outras Propriedades da Fila M/M/1

- Do mesmo modo, pode ser mostrado que a **CDF do tempo de espera** é dada por:

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

- Esta é uma distribuição exponencial truncada. O seu **Posto percentil *q*** é dado por:

$$w_q = \frac{1}{\mu(1-\rho)} \ln\left(\frac{100\rho}{100-q}\right)$$

- Esta fórmula se aplica apenas se q for maior do que 100(1-ρ). Todos os postos percentis mais baixos são 0

Outras Propriedades da Fila M/M/1

- **Número médio de jobs na fila**:

$$E[n_q] = \sum_{n=1}^{\infty} (n-1) p_n = \sum_{n=1}^{\infty} (n-1)(1-\rho)\rho^n = \frac{\rho^2}{1-\rho}$$

- O servidor é dito **ocioso** quando não houver nenhum job no sistema; em todos os demais momentos ele é dito **ocupado**

- O intervalo de tempo entre dois intervalos ociosos sucessivos é denominado de **período ocupado**

Exemplo 31.1

- Medições efetuadas em um roteador indicaram que os pacotes chegam a uma taxa média de 125 pacotes por segundo (pps) e o roteador leva aproximadamente 2 ms para encaminhá-los.

  a) Analise o roteador segundo um modelo M/M/1.

  b) Qual a probabilidade de estouro do buffer se o roteador tiver apenas 13 buffers?

Exemplos

Planejamento de Capacidade para Serviços na Web
Menascé & Almeida, 2003

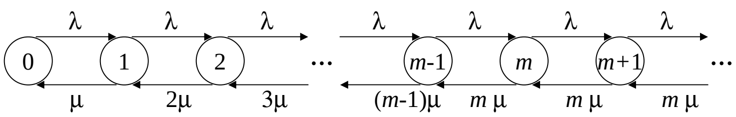
- **Exemplo 8.1:** As requisições chegam a um servidor web a uma taxa de de 30 requisições/segundo. Cada requisição gasta 0,02 segundo, em média, para ser processada. Qual a fração de tempo em que k (k = 0, 1, ...) requisições se encontram no servidor web? Qual o número médio de requisições no servidor?

- **Exemplo 8.2:** Considere o exemplo 8.1 acima. Responda:
  a) Qual o tempo médio de resposta no servidor?
  b) Qual seria o tempo médio de resposta se o servidor fosse substituído por outro duas vezes mais rápido?
  c) Qual seria o tempo de resposta se a taxa de chegada dobrasse quando o servidor fosse duas vezes mais rápido?

Fila M/M/m

Fila M/M/m

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



- Modelada como um processo de nascimento e morte onde:

$$\lambda_n = \lambda, \qquad n = 0,1,2,\ldots,\infty$$

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

Fila M/M/m

- Utilizando o Teorema 31.1, obtemos:

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

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

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

Fila M/M/m

- A probabilidade de haver 0 jobs no sistema é calculada pela relação:

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

- Que resulta em:

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

ou

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

Fórmula C de Erlang

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

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

- Para $m=1$ (servidor único), $C(m, \rho)$ é dada por ρ.

Número médio de jobs em espera

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

Número médio de jobs em atendimento

$$E[n_s] = \sum_{n=1}^{m-1} np_n + \sum_{n=m}^{\infty} mp_n$$
$$= 1p_0 \frac{(m\rho)}{1!} + 2\,p_0 \frac{(m\rho)^2}{2!} + \cdots + (m-1)\,p_0 \frac{(m\rho)^{m-1}}{(m-1)!}$$
$$+ m(p_m + p_{m+1} + p_{m+2} + \cdots +)$$
$$= m\rho\left(p_0 + p_0 \frac{(m\rho)}{1!} + p_0 \frac{(m\rho)^2}{2!} + \cdots + p_0 \frac{(m\rho)^{m-2}}{(m-2)!} \right) + mC(m,\rho)$$
$$= m\rho(p_0 + p_1 + p_2 + \cdots + p_{m-2}) + mC(m,\rho)$$
$$= m\rho[1 - p_{m-1} - C(m,\rho)] + mC(m,\rho)$$
$$= m\rho - m\rho p_{m-1} + mC(m,\rho)(1-\rho)$$
$$= m\rho \qquad \text{dado que } mC(m,\rho)(1-\rho) = mp_m = m\rho p_{m-1}$$

Número de Jobs no Sistema

- Número médio de jobs no sistema:

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

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

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

Utilização da Fila M/M/m

- Se observarmos o sistema por um tempo longo (p.ex., T segundos),
  - o número total de jobs chegando e recebendo serviço será λT
  - O tempo total ocupado dos *m* servidores para atender esses jobs será $\lambda T/\mu$
  - A **utilização** de cada servidor será dada por:

$$U = \frac{\text{tempo ocupado por servidor}}{\text{tempo total}} = \frac{(\lambda T / \mu)/m}{T} = \frac{\lambda}{m\mu} = \rho$$

  - Portanto, cada novo servidor acrescentado ao sistema diminui proporcionalmente sua utilização

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:

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

FDP do Tempo de Resposta

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

- 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

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:

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

Exemplo 31.2

- Estudantes chegam a um laboratório de computação de acordo com Poisson a uma taxa média de 10 por 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.

- Analise o laboratório usando um modelo de filas M/M/m.

Exemplo 31.3

- 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.
- Isso é viável? Se for, quantos computadores seriam necessários?

- *Sugestão: analise o sistema com m=6,7... computadores, mantendo as mesmas taxas de chegada e atendimento de λ=0,167 e μ=0,05, respectivamente.*

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

- Com m servidores, qual a melhor alternativa?
  - manter **filas separadas** para cada servidor, ou
  - 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 λ/m
  - uma fila M/M/m com taxa de chegada λ
- Vamos verificar que uma única fila é melhor quando os jobs são homogêneos

Exemplo 31.4

- No exemplo 31.2, 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, λ=0,167/5=0,0333 e μ=0,05, temos:

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

$$E[r] = \frac{1/\mu}{1-\rho} = \frac{1/0{,}05}{1-0{,}067} = 60$$

$$\text{Var}[r] = \frac{1/\mu^2}{(1-\rho)^2} = \frac{1/0{,}05^2}{(1-0{,}067)^2} = 3600$$

M/M/5
24
479

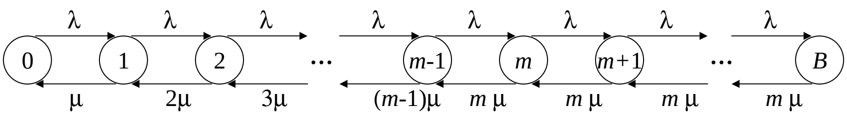
Fila M/M/∞

- Número infinito de servidores.
- É um caso especial das filas M/M/m
- Os jobs nunca têm que esperar, pois há sempre um servidor disponível
- O tempo de resposta (R) é igual ao tempo de serviço (S)
- 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

Fila M/M/m/B

Fila M/M/m/B (buffers finitos)

- Número de buffers B ($\geq m$) é finito
- Quando todos os buffers estiverem ocupados, novas chegadas serão perdidas:



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

$$\lambda_n = \lambda, \qquad n = 0,1,2,\ldots,B-1$$

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

Fila M/M/m/B

- Utilizando o Teorema 31.1, obtém-se:

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

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

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

Fila M/M/m/B

- A probabilidade de haver 0 jobs no sistema é calculada pela relação:

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

- Que resulta em:

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

- ou

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

Número Médio de Jobs

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

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

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

Taxa Efetiva de Chegadas

- Todas as chegadas que ocorrem enquanto o sistema está cheio *(n=B)* são perdidas
- A taxa dos jobs que efetivamente entram no sistema é dada por:

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

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

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:

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

Utilização da Fila M/M/m/B

- Se observarmos o sistema por um tempo longo, por exemplo, T segundos,
  - o número total de jobs chegando e recebendo serviço será $\lambda' T$
  - O tempo total ocupado dos m servidores para atender esses jobs será $\lambda' T/\mu$
  - 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

Fórmula de Perdas de Erlang

- Em um sistema M/M/m/m, o número de buffers 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 *1* servidor, caracterizando um sistema M/M/1/B

Exemplo 31.5

- Considere novamente o roteador do Exemplo 31.1. Analise o roteador assumindo que este possua apenas 2 buffers.
- A taxa média de chegadas e a taxa média de serviço, são, como antes, 125 pps e 500 pps, respectivamente.

Outros Sistemas de Filas

- G/M/1
- M/G/1 (M/D/1, caso especial)
- G/G/1
- G/G/m