PARTE III
TEORIA DE FILA

III.1 INTRODUÇÃO

Fila = linha de espera

Teoria de Fila:

- base matemática para a maioria dos modelos de sistemas de computação e de redes de comunicações.
- estudo do fenômeno da linha de espera.
- potente abstração para modelar sistemas que consistem de uma coleção de recursos de serviços e uma população de clientes.

Teoria de Fila:

- usada para estimar o desempenho de redes de computadores e de comunicações.

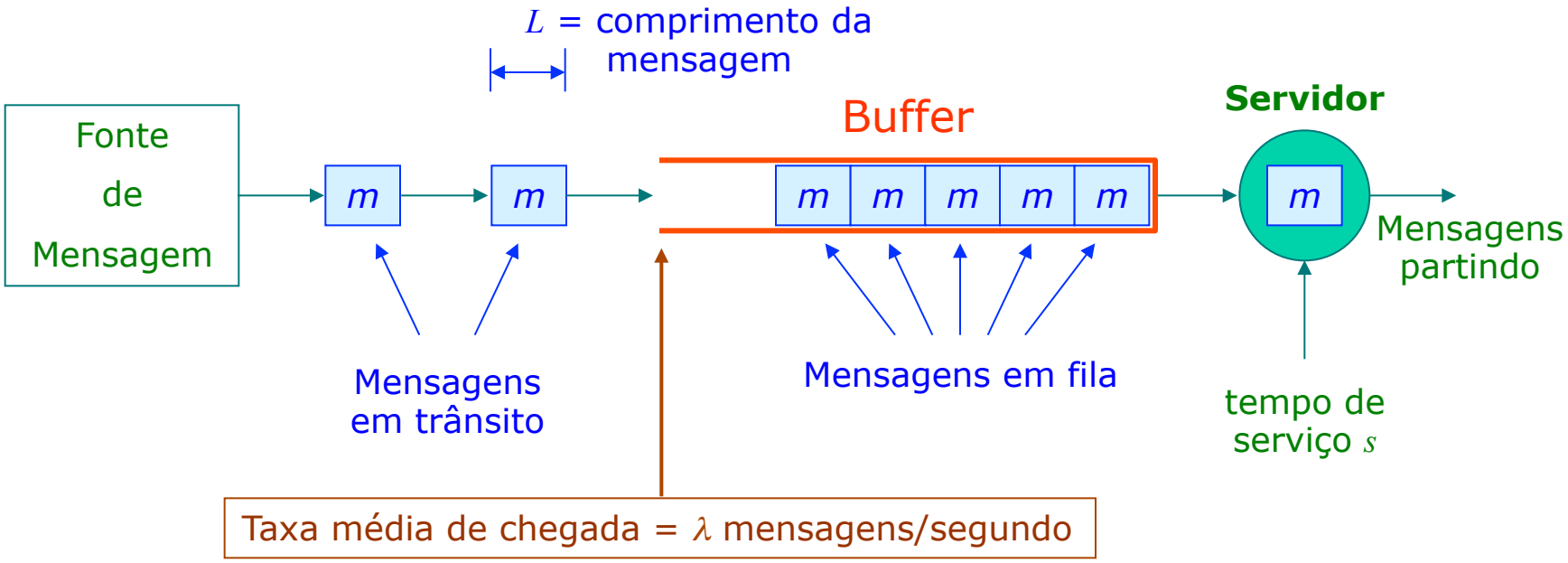
- Seu propósito é analisar a disputa por recursos e determinar seu efeito no fluxo de trabalho através de um sistema.

FSP, 08/06/2010

III.2 Descrição e Definições

III.2.1 Modelos de Filas

Fila com servidor único (Single-Server Queue):



As mensagens (*jobs*) que chegam ao buffer são armazenadas em fila e esperam pelo serviço de um único elemento de processamento (servidor único).

As mensagens que chegam ao buffer podem vir de um grupo de fontes que são diretamente conectadas ao nó (fila de mensagens) ou elas podem vir de uma linha externa que é conectada a outro nó.

A fonte de mensagem pode ser finita ou infinita.

Um sistema de fontes finitas não pode ter uma fila de serviço arbitrariamente longa, mas quanto maior for o número de fontes de mensagens maior será a taxa de chegada de mensagem.

Em um sistema de fontes infinitas, o comprimento da fila de serviço é ilimitado, e a taxa de chegada de mensagem não é afetada pelo número de fontes.

Mensagens chegam no buffer a uma taxa de λ mensagens/segundo.

Mensagens possuem um comprimento de X unidades de dados (bits, bytes, caracteres, etc.)

Assumimos: todas as mensagens possuem mesmo comprimento L.

Funções em um nó são caracterizadas pela taxa de serviço e pela disciplina da fila.

Taxa de serviço = número de jobs deixando o nó por unidade de tempo de serviço.

Taxa de serviço pode ser dependente da carga, isto é, pode depender do comprimento da fila.

A disciplina da fila = regra utilizada para determinar a ordem na qual os *jobs* enfileirados recebem o serviço.

Exemplo: Primeiro a chegar primeiro a ser servido (FCFS – first come first served): fila de bancos, supermercados, etc.

Assumiremos que as mensagens são processadas segundo a regra FCFS com taxa de C unidades de dados por segundo (capacidade).

Se uma mensagem chega e existem n mensagens a sua frente no buffer, então o tempo total T para processar essa mensagem consiste no tempo w gasto esperando na fila + o tempo de processamento s:

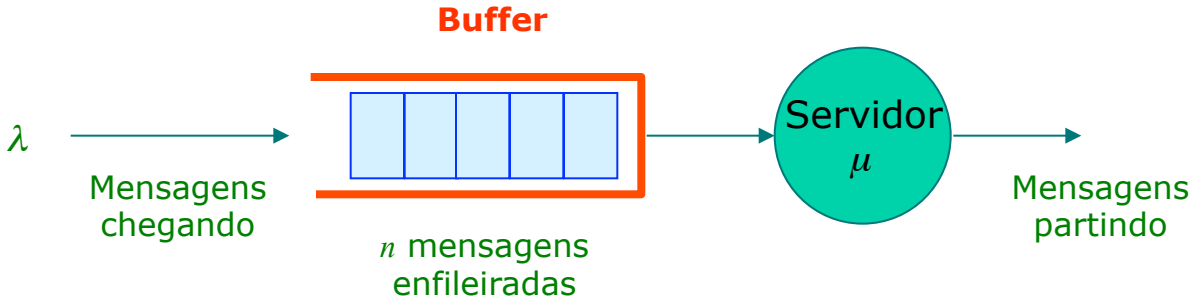
$$T = s + w = \frac{L}{C} + \frac{nL}{C} = \frac{1+n}{\mu}$$

onde $\mu = C/L$ é a taxa de serviço [mensagens/segundo].

T = tempo total de espera ou tempo de atraso.

= variável aleatória, pois o estado n do buffer varia com o tempo.

Modelo simples de uma fila de servidor único:



III.2.2 Estatística de Poisson

Sistema de fila:

- chegada da mensagem = variável aleatória
- tempo de serviço = variável aleatória

Processo de chegada de mensagens $\Rightarrow$ estatística de Poisson.

Estatística de Poisson é baseada em uma distribuição discreta de eventos.

Sistema possui um grande número de clientes independentes.

Definição: A probabilidade $P_n(t)$ de exatamente n clientes chegarem durante o intervalo de tempo t é dada por:

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

onde λ = taxa média de chegada e $n = 0, 1, 2, \ldots$

Se $N = \{n \mid n = 0, 1, 2, \ldots\}$ é o conjunto do possível número de mensagens que chegam a um nó, onde a probabilidade de n chegadas é dada pela equação acima, então o número esperado ou a média de chegadas de mensagens no intervalo de tempo T é dado por:

$$E[N] = \sum_{n=0}^{\infty} nP_n(t) = \lambda T$$

Processo de Poisson: estatística do processo de chegada de mensagens.

Assumimos que mensagens entram no sistema de fila nos instantes t_0, t_1, t_2, ..., t_n, onde $t_0 < t_1 < t_2 < \ldots < t_n$.

$\tau_n = t_n - t_{n-1}$ intervalo entre chegadas ($n > 0$) $\Rightarrow$ formam uma sequência de variáveis aleatórias independentes e identicamente distribuídas.

τ = um intervalo arbitrário entre chegadas com função densidade de probabilidade $a(t)$.

Pode-se provar que chegadas de Poisson geram uma densidade de probabilidade entre chegadas exponencial.

Prova:

Seja Δt um intervalo de tempo infinitesimal.

$a(t)\Delta t$ = probabilidade de que a próxima chegada ocorra no mínimo em t segundos mas não mais que $t + \Delta t$ segundos.

Ou seja, isso é a probabilidade $P_0(t)$ de nenhuma chegada para o tempo t multiplicada pela probabilidade $P_1(\Delta t)$ de uma chegada no tempo infinitesimal Δt.

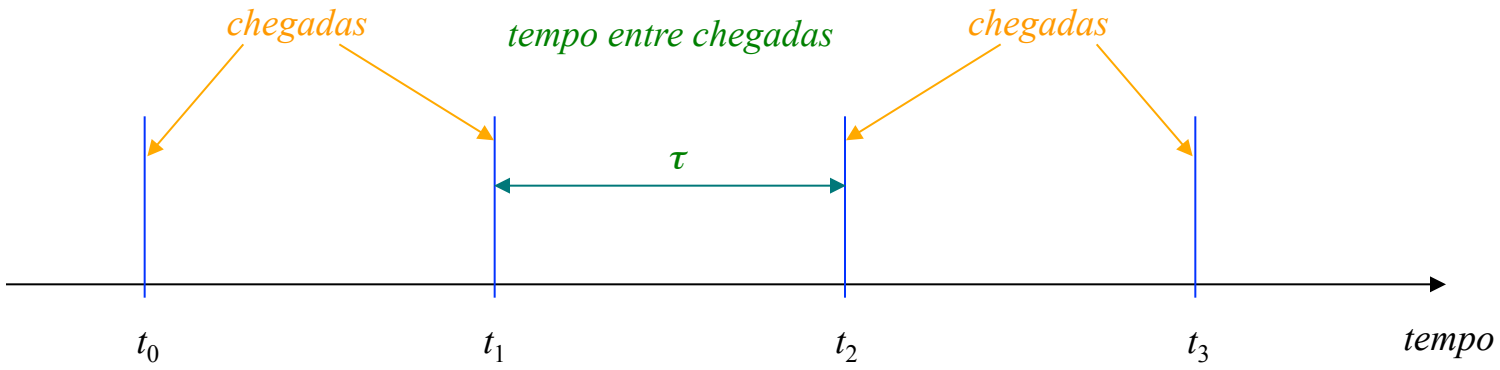
$$a(t)\Delta t = P_0(t)P_1(\Delta t)$$

onde $P_0(t) = \dfrac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$ $\qquad P_1(\Delta t) = \dfrac{(\lambda \Delta t)^1 e^{-\lambda \Delta t}}{1!} = \lambda \Delta t e^{-\lambda \Delta t}$

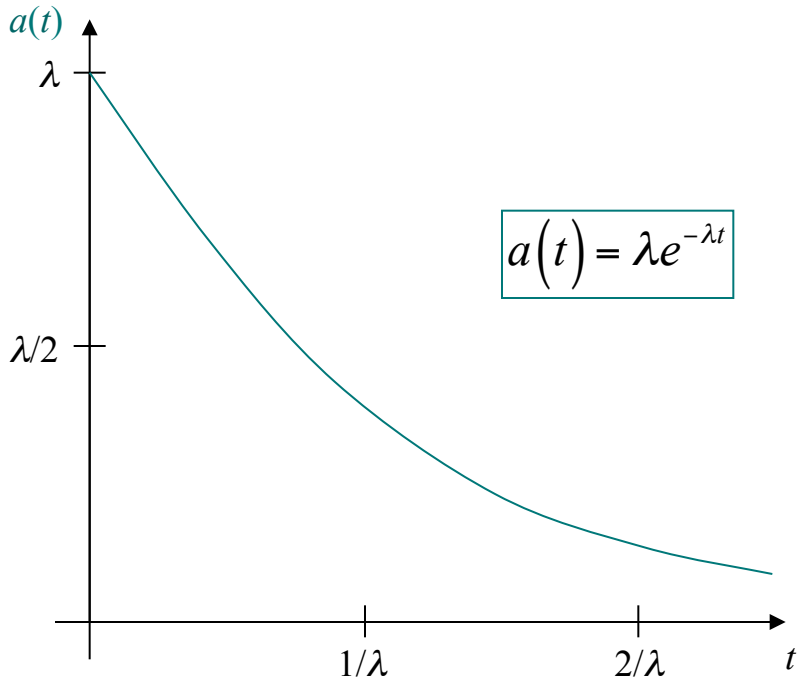
fazendo $\Delta t \to 0$, obtemos

$$\lim_{\Delta t \to 0} a(t)\Delta t = \lim_{\Delta t \to 0} e^{-\lambda t} \cdot \lambda \Delta t e^{-\lambda \Delta t} = \lambda e^{-\lambda t} dt \quad \Rightarrow \quad a(t)dt = \lambda e^{-\lambda t} dt$$

Chegadas de Poisson:



Tempo distribuído exponencialmente entre chegadas:



Assim, o tempo τ entre chegadas é uma variável aleatória exponencialmente distribuída e contínua $\Rightarrow$ processo de chegada de Poisson possui tempos de chegada exponenciais.

Tempo de serviço:

- tempo para completar a transmissão de uma mensagem,
- relacionado ao tamanho da mensagem.

Mensagens em uma fila são processadas pelo servidor a uma taxa fixa de C unidades de dados por segundo.

Mensagens variam em comprimento de maneira aleatória $\Rightarrow$ tempo de serviço dessas mensagens varia!

Assumimos que as mensagens são exponencialmente distribuídas em comprimento com comprimento médio igual a L.

Então, a mensagem será transmitida ou processada em L/C segundos.

Distribuição do tempo de serviço:

$$S(t) = \left(\frac{C}{L}\right) e^{-Ct/L}$$

Valor esperado de $S(t)$, isto é, o tempo médio de transmissão ou tempo médio de serviço da mensagem:

$$E\left[S(t)\right] = \frac{L}{C} = \frac{1}{\mu}$$

$\mu = C/L$ é a taxa de serviço [mensagens/segundo].

A distribuição do tempo de serviço $S(t)$ possui a mesma forma de $a(t)$, então, ela obedece a estatística de Poisson.

III.2.3 Notação (D. G. Kendall)

Forma mais geral: A/B/c/K/N/Z

A: distribuição de tempo entre chegadas

B: disciplina de serviços

c: número de servidores

K: capacidade do sistema (máximo comprimento de fila ou tamanho de buffer)

N: número de usuários potenciais em uma dada população fonte

Z: disciplina de enfileiramento

Quando não existe limite no comprimento da linha de espera (fila), o número de fontes é infinito e as mensagens (*jobs*) são aceitas na base do “primeiro a chegar primeiro a ser servido” (FCFS).

Assim, a notação mais curta A/B/c é geralmente utilizada.

Símbolos usados para A e B:

GI: tempos entre chegadas independentes gerais.

G: distribuição de tempo de serviço geral.

M: distribuição de tempo de serviço ou de entre chegadas exponencial (Markov).

D: distribuição de tempo de serviço ou de entre chegadas determinística (constante).

Exemplos:

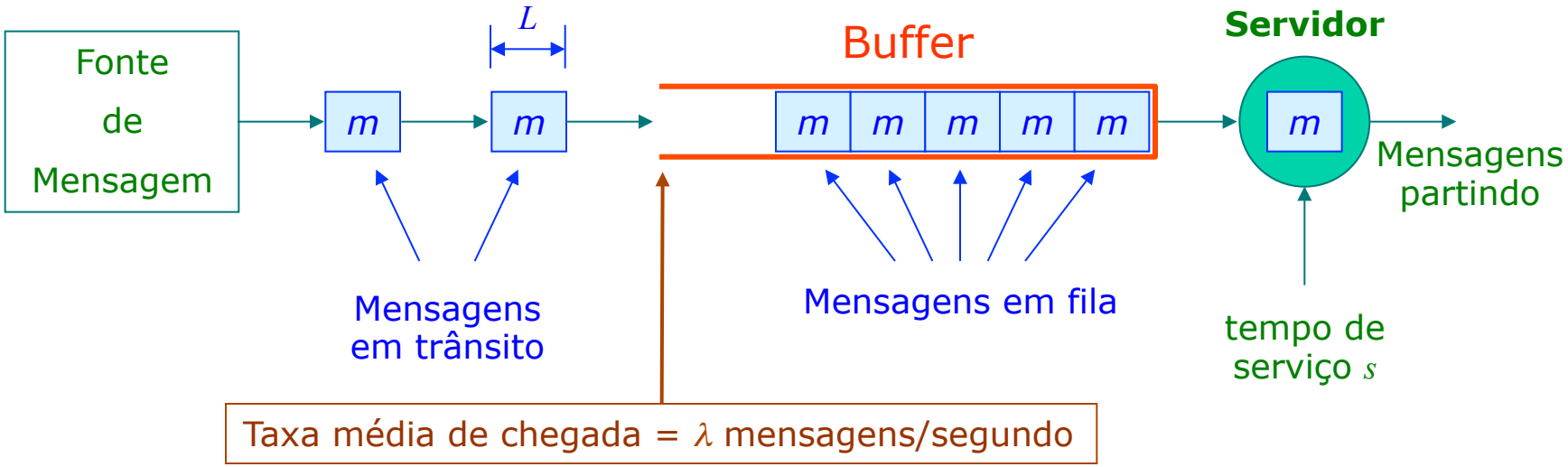
Fila M/G/1: Fila possui chegadas de Poisson, uma distribuição de serviço geral (distribuição de tempo de serviço pouco definido) e um servidor.

Fila M/D/1: Fila possui chegadas de Poisson, um tempo de serviço constante ou fixo e um servidor.

Fila M/M/1: Mensagens chegam com distribuição de Poisson, entram em um buffer infinito com uma distribuição de tempo de serviço exponencial e são manipuladas por um servidor na base primeiro a chegar primeiro a ser servido (FCFS).

Fila M/M/1: Modelo amplamente usado devido a suas distribuições de probabilidades que descrevem o processo de entrada e o processo de serviço de uma forma matemática simples.

Oferece um modelo mais realístico para sistemas de fila reais, pois os padrões de chegada de cliente em sistemas reais seguem uma distribuição de probabilidade de Poisson.



III.2.4 Relações entre Variáveis Aleatórias

Relações básicas entre variáveis aleatórias que descrevem o número de mensagens em várias partes do sistema e as variáveis aleatórias que descrevem tempo (tempo de fila, tempo de serviço, etc.)

$N(t)$ = número de mensagens em um sistema no tempo t.

$N_q(t)$ = número de mensagens na fila no tempo t.

$N_s(t)$ = número de mensagens sendo processadas no tempo t.

$$N(t) = N_q(t) + N_s(t)$$

Início de operação $\Rightarrow$ transiente $\Rightarrow$ equação difícil de resolver.

Transiente:

Ocorre devido ao sistema no instante t poder possuir um certo número de mensagens na fila e um certo número de mensagens sendo servidas (condições iniciais).

Após o início da operação do sistema, a influência das condições iniciais vão diminuindo até desaparecer.

Assim, o número de mensagens no sistema e na fila são independentes do tempo.

O sistema é dito em equilíbrio ou em regime permanente (*steady state*).

Equilíbrio:

$$\tilde{N} = \tilde{N}_q + \tilde{N}_s$$

onde estas variáveis aleatórias são independentes e descritas por suas distribuições de probabilidades.

Média:

$$E\left[\tilde{N}\right] = E\left[\tilde{N}_q\right] + E\left[\tilde{N}_s\right]$$

ou

$$N = N_q + N_s$$

N = número médio do total de mensagens no sistema (em espera e sendo servidas).

N_q = número médio de mensagens na fila.

N_s = número médio de mensagens sendo servidas.

Relações similares de variáveis aleatórias temporais:

w = tempo total que uma mensagem gasta no sistema (em espera + servida).

q = tempo de espera na fila para uma mensagem.

s = tempo de serviço para uma mensagem.

$$w = q + s$$

Assim,

$$E[w] = E[q] + E[s]$$

No equilíbrio temos:

$$T = T_q + T_s$$

T = tempo médio que uma mensagem gasta no sistema (tempo de espera e de serviço).

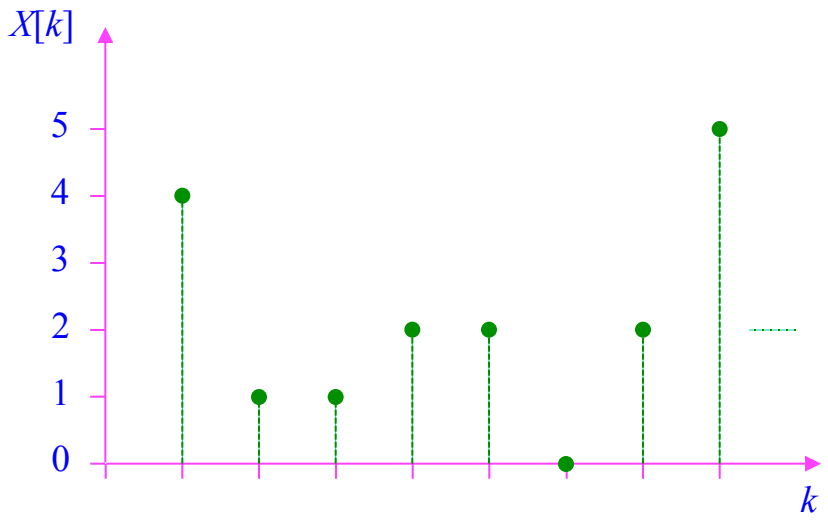
T_q = tempo médio de espera da mensagem na fila.

T_s = tempo médio para processar uma mensagem.

III.2.5 Cadeias de Markov Discretas no Tempo

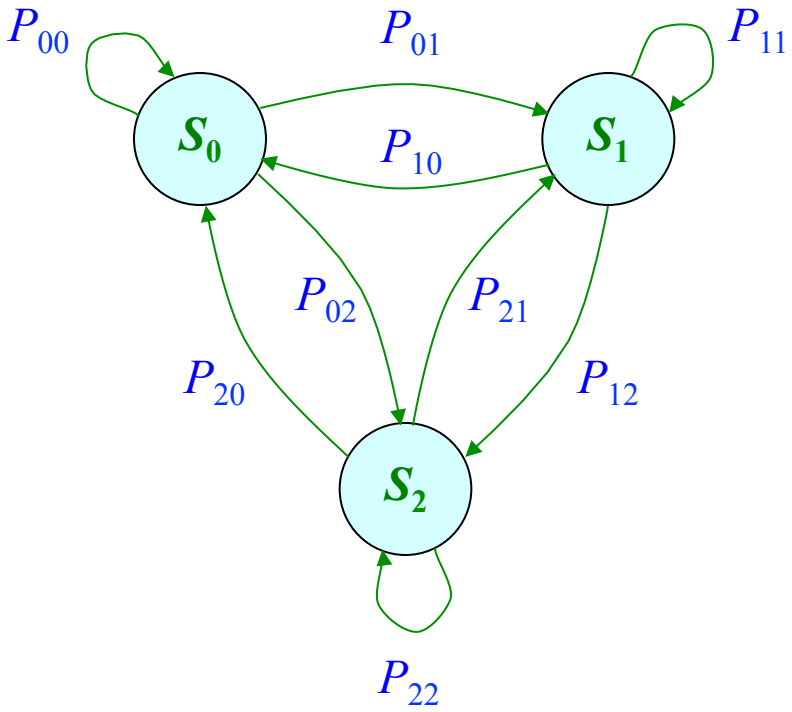
Cadeia de Markov com um conjunto finito de estados = processo aleatório que assume valores contidos em um conjunto discreto $\{0, 1, 2, 3, ..., q\text{-}1\}$.

Realização típica:



$X[k]$ = estado do processo no instante de tempo k.

Representação em diagrama de estados:

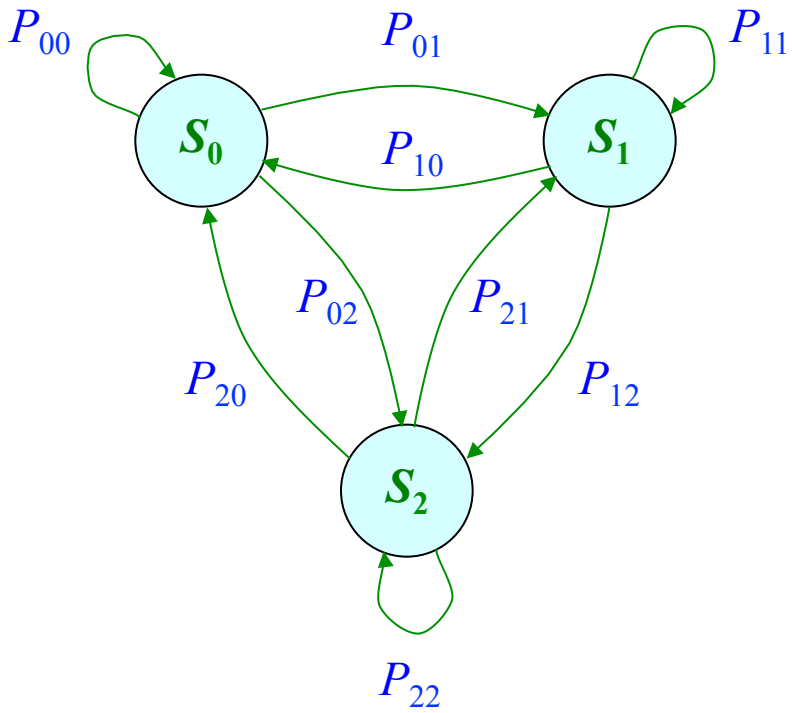


Processo de Markov é definido pelo conjunto de probabilidades:

$$p_i[k] = \Pr[X[k] = i] = \Pr[\text{ de estar no estado } S_i \text{ no instante } k]$$

$$P_{ij} = \Pr[X[k+1] = j \mid X[k] = i] = \Pr[\text{ de transição do estado } S_i \text{ para o estado } S_j]$$

Exemplo:



Probabilidade de estar no estado i no instante k+1: $\mathbf{p}[k+1] = \mathbf{P}^T \mathbf{p}[k]$

$$p_0[k+1] = P_{00}p_0[k] + P_{10}p_1[k] + P_{20}p_2[k]$$
$$p_1[k+1] = P_{01}p_0[k] + P_{11}p_1[k] + P_{21}p_2[k]$$
$$p_2[k+1] = P_{02}p_0[k] + P_{12}p_1[k] + P_{22}p_2[k]$$

ou

$$\begin{bmatrix} p_0[k+1] \\ p_1[k+1] \\ p_2[k+1] \end{bmatrix} = \begin{bmatrix} P_{00} & P_{10} & P_{20} \\ P_{01} & P_{11} & P_{21} \\ P_{02} & P_{12} & P_{22} \end{bmatrix} = \begin{bmatrix} p_0[k] \\ p_1[k] \\ p_2[k] \end{bmatrix}$$

III.3 Filas M/M/1

III.3.1 Transições de estado

Sistema M/M/1 de fila em equilíbrio:

> A probabilidade de encontrar o sistema em um dado estado n não muda com o tempo.

Se um cliente chega, o estado muda de n para $n+1$, enquanto que quando o cliente foi servido e parte, o estado muda de n para $n-1$.

Para o sistema estar em equilíbrio estes dois processos devem ocorrer com mesma taxa.

Este princípio é conhecido como princípio de equilíbrio detalhado.

Análise da probabilidade de estado de um sistema de fila M/M/1:

Assumimos que as transições são somente entre estados adjacentes.

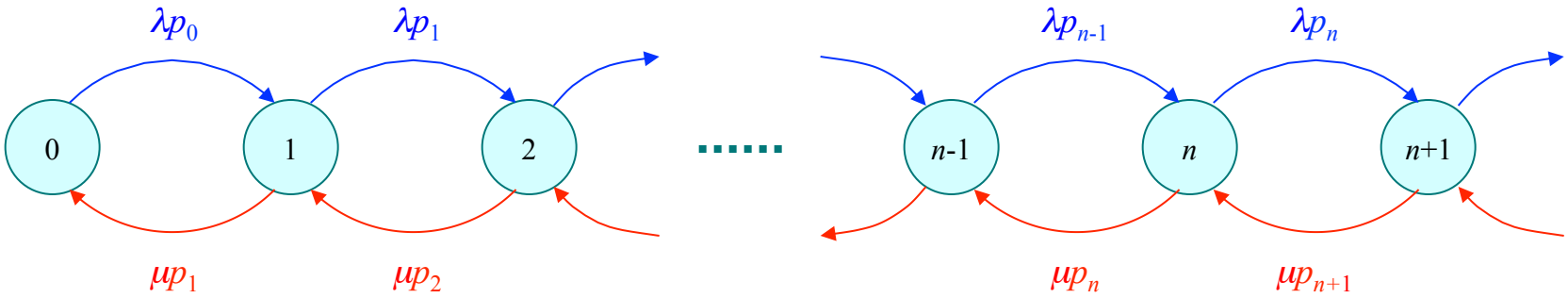
Se o sistema está no estado n (existem n clientes ou jobs no sistema consistindo de uma fila mais um servidor) e um novo cliente chega, o sistema se move para o estado $n+1$.

Quando um cliente é servido e parte, o sistema se move para o estado $n-1$.

Sistemas onde só ocorrem transição entre estados adjacentes são chamados de sistemas nascimento e morte (*birth-death systems*).

Chegada de um cliente = nascimento. Partida de um cliente = morte.

Diagrama de estado para fila M/M/1:



λ = taxa média de chegada de clientes por segundo.

λp_n = número médio de transições por segundo do estado n para o estado n+1.

p_n = probabilidade de equilíbrio de que há exatos n clientes no sistema (fila + servidor).

μ = taxa média de processamento de clientes por segundo.

μp_{n+1} = número médio de transições por segundo do estado $n+1$ para o estado n.

III.3.2 Tamanho da Fila

Sistema M/M/1 de fila em equilíbrio $\Rightarrow$ taxa de chegada de mensagens é igual taxa de partida de mensagens.

Então, o fluxo de mensagens que flui para o estado n é igual ao de mensagens que flui para fora dele.

Desejamos encontrar: $p_n(t)$ = probabilidade de n mensagens estarem presentes no buffer no tempo t.

Assumimos $\lambda_n = \lambda$ para qualquer estado e que $\mu_n = \mu$ para estados $n > 0$.

Para $n = 0$, temos $\mu_0 = 0$.

Taxa de chegada e taxa de serviço de Poisson.

No intervalo $(t, t + \Delta t)$, as seguintes transições podem ocorrer:

1. Uma mensagem entra no estado n vinda do estado $n - 1$ com taxa de fluxo = λp_{n-1}.
2. Uma mensagem entra no estado n vinda do estado $n + 1$ com taxa de fluxo = μp_{n+1}.
3. Uma mensagem deixa o estado n para o estado $n - 1$ com taxa de fluxo = μp_n.
4. Uma mensagem deixa o estado n para o estado $n + 1$ com taxa de fluxo = λp_n.

Então, temos:

Fluxo para o estado $n = \lambda p_{n-1} + \mu p_{n+1}$

Fluxo saindo do estado $n = \lambda p_n + \mu p_n$

Assim, para $n > 0$, temos:

$$\lambda p_{n-1} + \mu p_{n+1} = (\lambda + \mu) p_n$$

Condições de contorno:

- Transição do estado 0 (estado vazio) para o estado 1: λp_0
- Transição do estado 1 para o estado 0: μp_1

Assim, os dois fluxos tem que ser iguais:

$$\lambda p_0 = \mu p_1$$

ou

$$p_1 = \frac{\lambda}{\mu} p_0 = \rho p_0$$

$\rho = \lambda/\mu$ é a intensidade de tráfego, determina o número mínimo de servidores necessário para manter a chegada de mensagens (unidade = Erlang).

Exemplo: Sistema de fila com tempo médio entre chegadas $1/\lambda = 10$ s.
Tempo médio de serviço $1/\mu = 5$ s.

Razão de intensidade de tráfego: $\rho = \frac{\lambda}{\mu} = \frac{1/10}{1/5} = 0{,}5$ Erlangs

Se tempo médio de serviço $1/\mu = 15$ s:

Então, $\rho = 1{,}5$ Erlangs

O que indica que as mensagens chegam mais rápidas que elas podem ser processadas.

Expressão geral:

Tomando $\lambda p_{n-1} + \mu p_{n+1} = (\lambda + \mu) p_n$ para $n = 1$, temos:

$$\lambda p_0 + \mu p_2 = (\lambda + \mu) p_1$$

como

$$p_1 = \frac{\lambda}{\mu} p_0 = \rho p_0$$

temos

$$p_2 = (\rho + 1) p_1 - \rho p_0 = \rho^2 p_0$$

Repetindo este processo recursivamente, obtemos:

$$\boxed{p_n = \rho^n p_0}$$

Mas sabemos que

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

$\rho = \lambda / \mu$

Como as probabilidades de estado devem decrescer com n, então devemos ter

$$\rho = \frac{\lambda}{\mu} < 1$$

O número médio de chegadas por unidade de tempo deve ser menor que a capacidade do sistema, isto é

$$\lambda < \mu$$

Caso contrário, o estado de regime permanente (*steady state*) não poderá ser alcançado.

Sabemos que

$$\sum_{n=0}^{\infty} p_n = 1 = p_0 \sum_{n=0}^{\infty} \rho^n = \frac{p_0}{1-\rho}$$

Relação utilizada para $x < 1$:

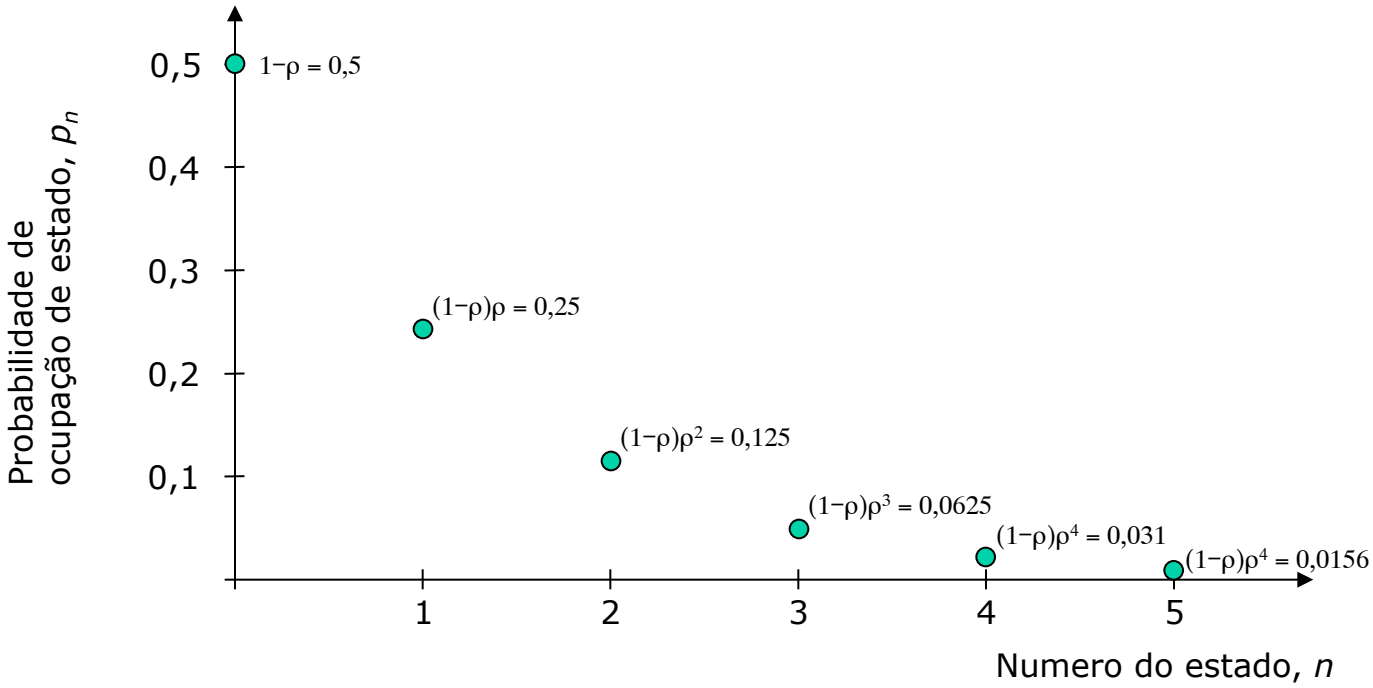
$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

Então, a probabilidade de ocupação dos vários estados da fila M/M/1 é:

$$p_n = (1-\rho)\rho^n$$

Note que devemos ter $\rho < 1$ e que quando a intensidade de tráfego ρ cresce, os estados de maior ordem se tornam relativamente mais prováveis.

Exemplo: $\rho = 0{,}5$



III.4 Fórmula de Little

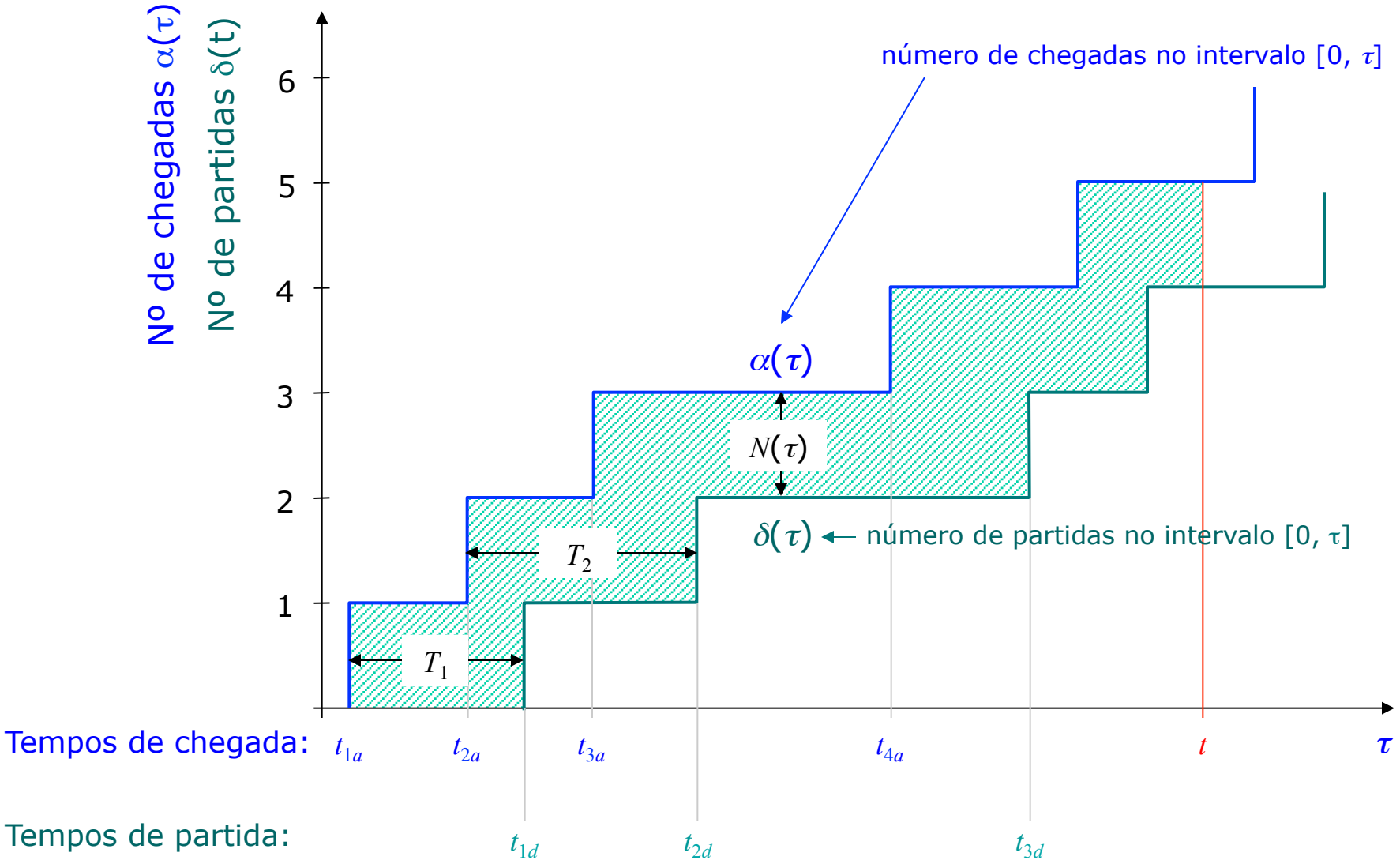
Parâmetros importantes em sistemas de fila:

- Número médio de clientes esperando na fila.
- Tempo médio que estes clientes gastam no sistema esperando pelo serviço.

Para um estado arbitrário existem 4 quantidades envolvidas: N, N_q, T e T_q.

A fórmula de Little permite determinar qualquer uma destas quantidades dada uma delas.

III.4.1 Obtenção Gráfica da Fórmula de Little



Mensagens são processadas pela ordem de chegada (FCFS).

$\alpha(\tau)$ = número de mensagens chegadas no intervalo $[0, \tau]$.

$\delta(\tau)$ = número de mensagens que partem no intervalo $[0, \tau]$.

Exemplo: mensagem i chega no tempo t_{ia} e parte no tempo t_{id}.

Se o sistema está vazio no instante $\tau = 0$, o número de mensagens $N(t)$ no sistema no tempo t é:

$$N(t) = \alpha(t) - \delta(t)$$

O tempo total $\gamma(t)$ que todas as mensagens despenderam no sistema no instante t é a área cumulativa no intervalo de $[0, t]$:

$$\gamma(t) = \int_0^t N(\tau)\, d\tau \quad \text{ou} \quad \gamma(t) = \sum_{i=1}^{\delta(t)} T_i + \sum_{i=\delta(t)+1}^{\alpha(t)} (t - t_i)$$

T_i = tempo que a mensagem i gasta no sistema.

Igualando as expressões e dividindo por t, obtemos:

$$N_t = \lambda_t T_t$$

onde no intervalo $[0, t]$:

$$N_t = \frac{1}{t}\int_0^t N(\tau)\,d\tau$$

= nº médio de mensagens no sistema

$$T_t = \frac{1}{\alpha(t)}\left[\sum_{i=1}^{\delta(t)} T_i + \sum_{i=\delta(t)+1}^{\alpha(t)} (t - t_i)\right]$$

= tempo médio que uma mensagem gasta no sistema

$$\lambda_t = \frac{\alpha(t)}{t}$$

= taxa média de chegada de mensagem

Equação de Little obtida para o caso especial de serviço FCFS no intervalo finito $[0, t]$:

$$N_t = \lambda_t T_t$$

Para o caso de equilíbrio, $t \to \infty$, assumimos que N_t, λ_t e T_t permanecem finitos e tendem para seus valores de equilíbrio N, λ e T.

Fórmula de Little:

$$N = \lambda T$$

O número médio de clientes em um sistema de fila é igual a taxa média de chegada de clientes ao sistema multiplicada pelo tempo médio gasto no sistema.

De modo similar temos a relação:

$$N_q = \lambda T_q$$

O número médio de clientes em uma fila é igual a taxa média de chegada de clientes ao sistema multiplicada pelo tempo médio que o cliente gasta esperando na fila.

III.4.2 Aplicação em uma Fila M/M/1

Encontrar N, N_q, T e T_q em uma fila M/M/1.

Para $n > 0$, temos:

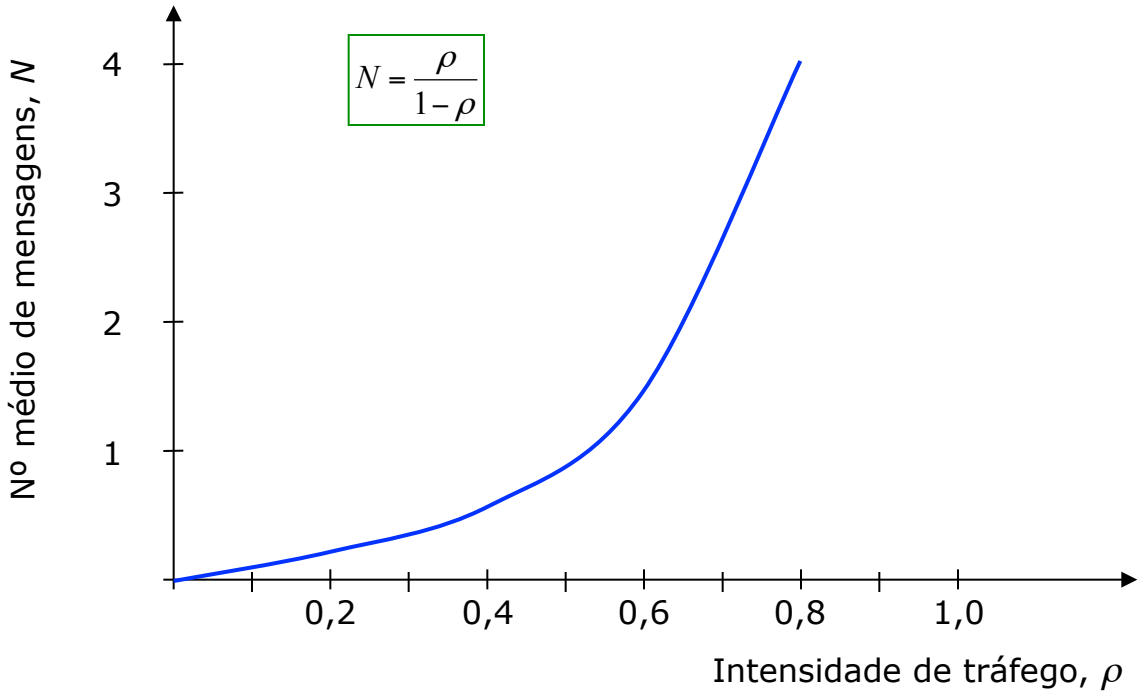
$$\lambda p_{n-1} + \mu p_{n+1} = (\lambda + \mu) p_n$$

que é fundamental para encontrar todas as quantidades estatísticas de interesse para uma fila.

Número médio N de mensagens no sistema:

$$N = E\left[\tilde{N}\right] = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\rho}$$

Comportamento do número esperado de mensagens no sistema:

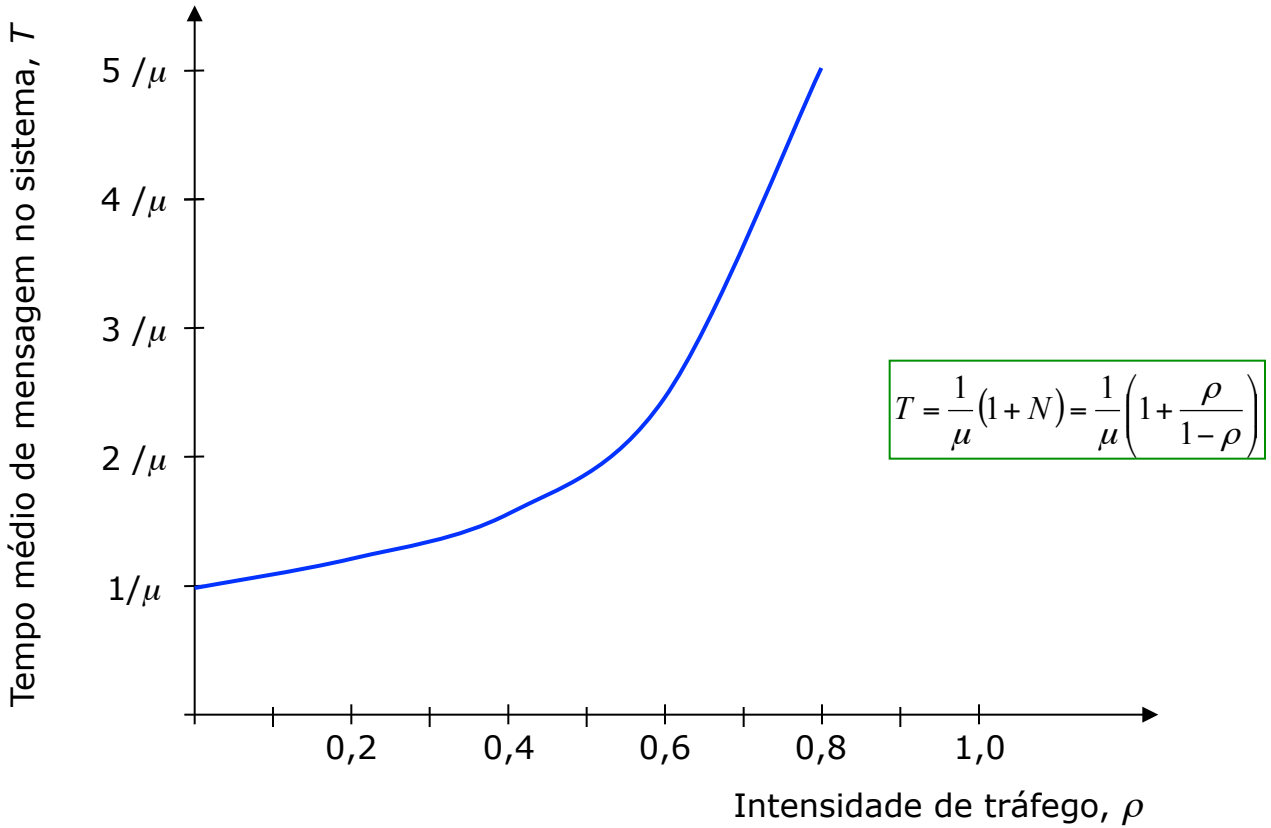


Para $\rho < 0{,}5 \Rightarrow N < 1$.

Para ρ grande $\Rightarrow$ nº de mensagens esperando cresce rapidamente.

Usando a fórmula de Little $N = \lambda T$, temos o tempo médio de espera no sistema:

$$T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu}(1+N)$$



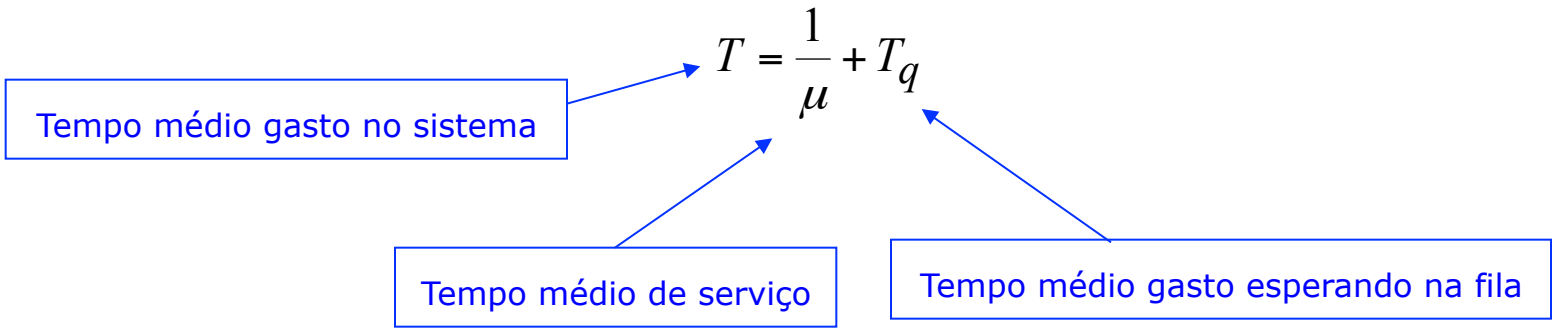
Note que o comportamento do tempo médio que uma mensagem gasta no sistema em função da intensidade de tráfego ρ é similar ao exibido pelo número médio de mensagens no sistema em função da intensidade de tráfego ρ.

Para $\rho = 0 \Rightarrow T = 1/\mu$ (tempo de serviço de uma mensagem) $\Rightarrow$ mensagem não tem que esperar numa fila e é, portanto, servida em $1/\mu$ segundos.

Para $\rho \rightarrow 1 \Rightarrow$ o número médio de mensagens no sistema e o tempo médio gasto no sistema aumentam drasticamente.

Assim, neste sistema de fila M/M/1 a penalidade paga é muito alta quando tentamos trabalhar perto a sua capacidade.

Analisando apenas a fila, temos:



Então,

$$T_q = T - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}$$

Aplicando a fórmula de Little $N_q = \lambda T_q$, obtemos o número médio de mensagens na fila:

$$N_q = \lambda \frac{\rho}{\mu(1-\rho)} = \frac{\rho^2}{1-\rho}$$

$$\rho = \frac{\lambda}{\mu}$$

Probabilidade de encontrar pelo menos n mensagens no sistema:

Usando $p_n = (1-\rho)\rho^n$, temos que a probabilidade de o número de mensagens em uma fila M/M/1 ser maior que um número N é dada por

$$P(n > N) = \sum_{n=N+1}^{\infty} p_n = (1-\rho) \sum_{n=N+1}^{\infty} \rho^n = \rho^{N+1}$$

Esta probabilidade é usada para calcular o tamanho do buffer.

Exemplo: $\rho = 0{,}5$

N	$P(n > N)$
1	0,250
2	0,063
5	0,016
10	$4{,}9\times10^{-4}$
15	$1{,}5\times10^{-5}$

A ocupação média do sistema é:

$$E[n] = \frac{\rho}{1-\rho} = \frac{0{,}5}{1-0{,}5} = 1$$

A chance da ocupação exceder a ocupação média do buffer em 10 vezes é menor que 10^{-3}.

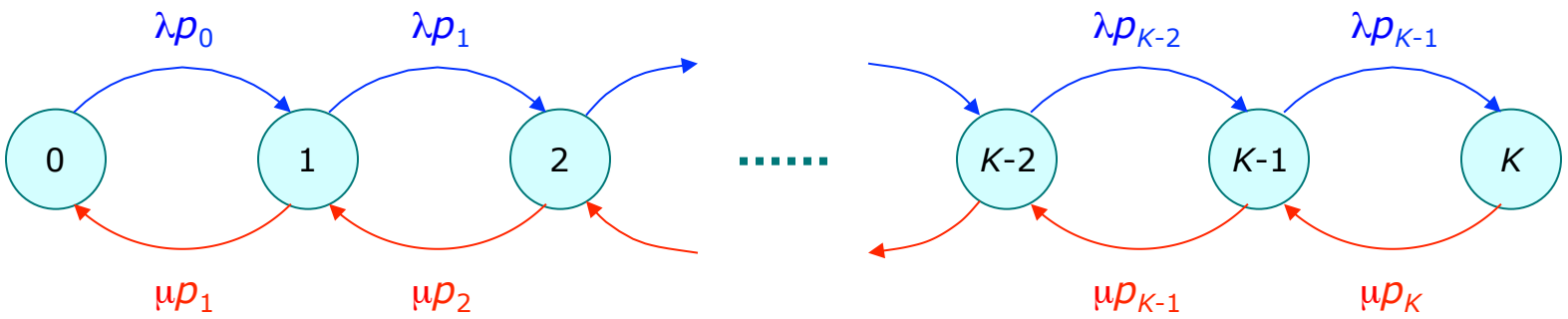
Um buffer que armazena 15 mensagens parecerá um buffer infinito.

III.5 FILAS M/M/1/K

Sistema de fila com número limitado de K mensagens que podem ser armazenadas, incluindo a que está sendo servida.

Qualquer mensagem extra que chegar, esta será descartada pelo sistema.

Diagrama de estado:



Assumindo que as mensagens chegam e são processadas a taxas constantes.

Coeficientes de taxa de transição:

$$\lambda_n = \begin{cases} \lambda & \text{para } n = 0, 1, 2, \ldots, K-1 \\ 0 & \text{para } n \geq K \end{cases}$$

$$\mu_n = \begin{cases} \mu & \text{para } n = 1, 2, \ldots, K \\ 0 & \text{para } n > K \end{cases}$$

Novas mensagens que chegam na fila serão geradas de acordo com um processo de Poisson mas só até atingir o valor de K mensagens.

Probabilidades de estados em regime (*steady state*) são dadas por:

$$p_n = p_0 \prod_{i=0}^{n-1} \frac{\lambda}{\mu} = \rho^n p_0 \qquad \text{para } n \le K$$

$$p_n = 0 \qquad \text{para } n > K$$

Resolvendo para p_0:

Probabilidades do conjunto finito de estados deve somar 1, então

$$\sum_{n=0}^{K} p_n = 1 = p_0 \sum_{n=0}^{K} \rho^n = p_0 \frac{1-\rho^{K+1}}{1-\rho} \quad \Rightarrow \quad \boxed{p_0 = \frac{1-\rho}{1-\rho^{K+1}}}$$

$$p_0 = \frac{1-\rho}{1-\rho^{K+1}}$$

$$p_n = \rho^n p_0 = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} \qquad \text{para } 0 \le n \le K$$

$$p_n = 0 \qquad \text{fora}$$

Note que para atingir o equilíbrio não é necessário que $\lambda < \mu$.

Se $\lambda = \mu$, então $\rho = 1$ e usando L´Hospital, temos:

$$\lim_{\rho \to 1} p_n = \lim_{\rho \to 1} \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} = \frac{1}{K+1} \qquad \text{para } 0 \le n \le K$$

Para $\rho^K << 1$:

$$p_n = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} \quad \Rightarrow \quad p_n \cong (1-\rho)\rho^n \qquad \text{(buffer infinito)}$$

Assim, a probabilidade do buffer estar cheio e de que as mensagens serão jogadas fora ou bloqueadas é simplesmente a probabilidade de existir K mensagens no sistema, ou seja:

$$p_K = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}$$

De modo similar à fila M/M/1, podemos aplicar a fórmula de Little para encontrar N, T, N_q e T_q.

Supondo $\lambda < \mu$, temos o número médio N de mensagens no sistema:

$$N = E\left[\tilde{N}\right] = \sum_{n=0}^{K} np_n$$

$$= \frac{\rho}{1-\rho} - \frac{\left(K+1\right)\rho^{K+1}}{1-\rho^{K+1}}$$

O número médio N_q de mensagens esperando na fila será:

$$N_q = N - N_s$$

Mas o número médio de mensagens sendo servidas é a probabilidade do sistema não estar vazio $(1 - p_0)$ vezes o número médio de mensagens que são servidas sob esta condição (que é igual a 1), ou seja,

$$N_s = E\left[\tilde{N}_s\right] = \left(1-p_0\right)\cdot 1 = 1-p_0$$

Se existe K mensagens no sistema M/M/1/K, então qualquer mensagem que chega não será autorizada a entrar no sistema até que uma outra seja processada e saia deste sistema.

Isto ocorre com probabilidade:

$$p_K = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}$$

A probabilidade de que uma mensagem possa entrar no sistema é igual a $(1 - p_K)$, tal que a taxa média λ_a de mensagens entrando no sistema é dada por:

$$\lambda_a = \lambda(1-p_K)$$

λ = taxa de chegada de mensagem.

Usando a fórmula de Little, obtemos o tempo médio para uma mensagem passar pelo sistema:

$$T = \frac{N}{\lambda_a}$$

e o tempo médio que uma mensagem gasta esperando na fila:

$$T_q = \frac{N_q}{\lambda_a}$$

Exemplo:

Fila M/M/1 com capacidade K.

Probabilidade de rejeição de mensagens é aproximadamente: $P[N = K]$

Comparação desta probabilidade aproximada com a probabilidade exata dada para o modelo M/M/1/*K*:

Para o sistema M/M/1:

$$P\left[N = K\right] = \left(1-\rho\right)\rho^{K}$$

Se $\rho < 1$, a probabilidade de rejeitar uma mensagem para o sistema M/M/1/*K* é:

$$P\left[N' = K\right] = \frac{\left(1-\rho\right)\rho^{K}}{1-\rho^{K+1}} = \left(1-\rho\right)\rho^{K}\left[1+\rho^{K+1}+\left(\rho^{K+1}\right)^{2}+\ldots\right]$$

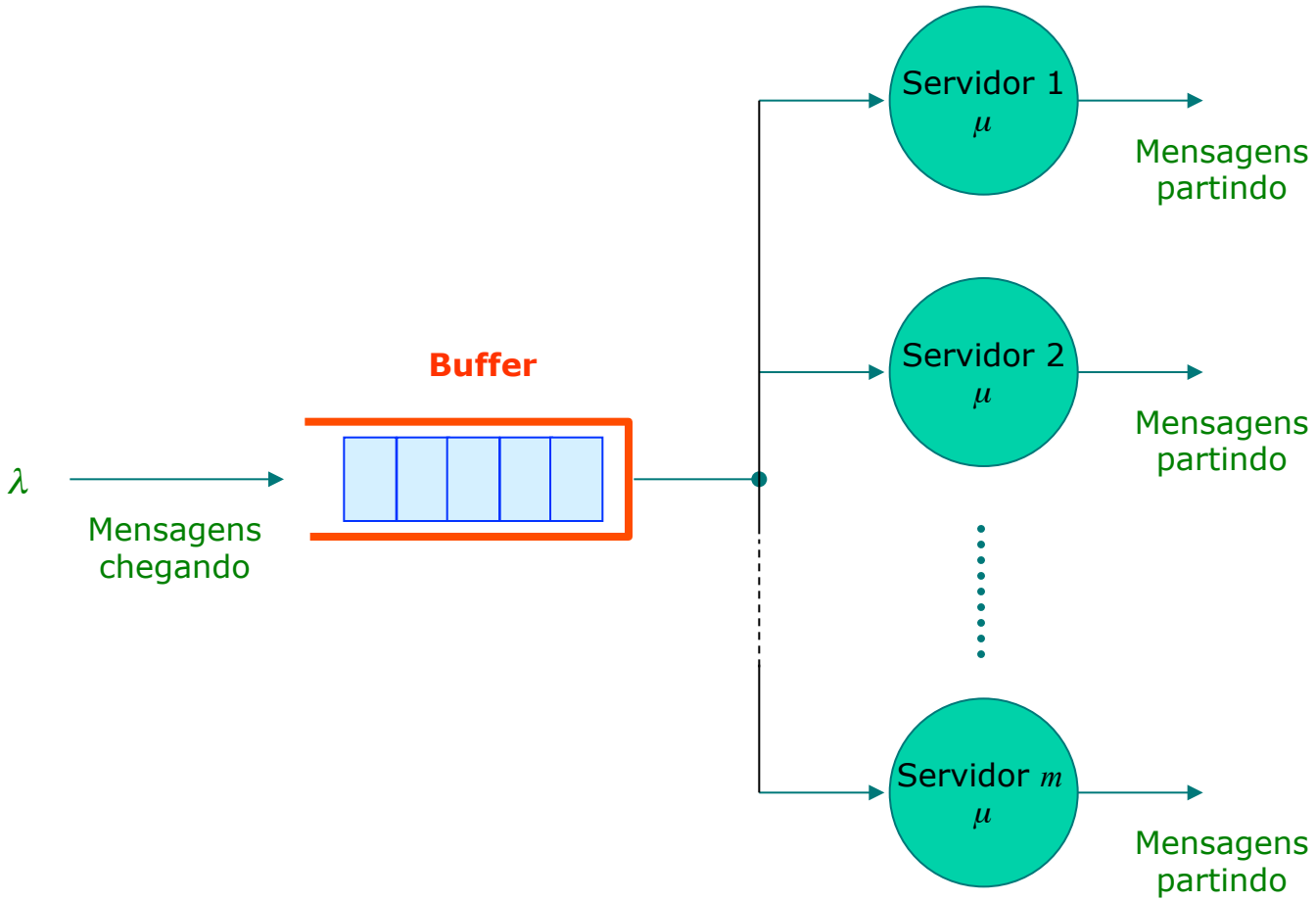
Se $\rho < 1$ e K grande:

$$P\left[N = K\right] \cong P\left[N' = K\right]$$

Se $\rho > 1$, a aproximação se quebra e fornece uma probabilidade negativa.

III.6 FILAS M/M/m

Neste caso o número de servidores é $m > 1$.



Assumimos:

- Todos os servidores possuem o mesmo destinatário.
- O sistema possui buffer infinito.
- Taxa de chegada λ_n para qualquer estado é igual a constante λ.
- Todos servidores são idênticos, cada um com capacidade de processamento C.

Taxas de serviço:

$$\mu_n = \min\left[n\mu, m\mu\right] = \begin{cases} n\mu & \text{para } 0 \le n \le m \\ m\mu & \text{para } n \ge m \end{cases}$$

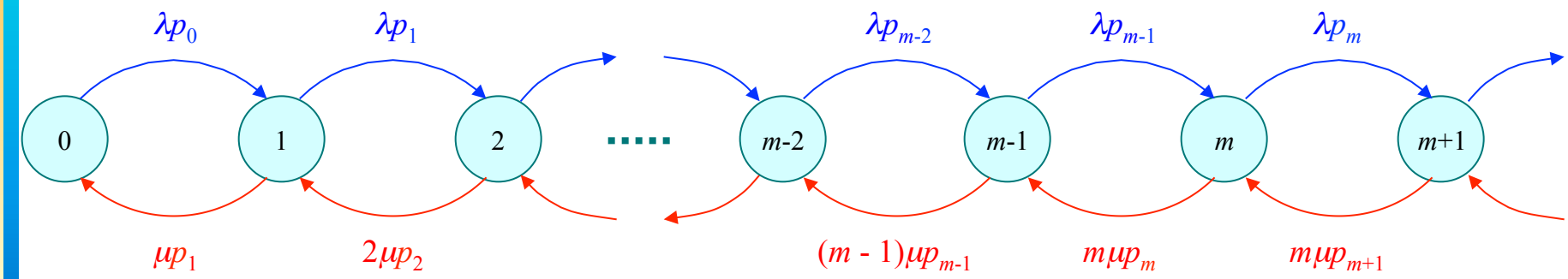
Definição: Fator de utilização do servidor, U:

$$U = \frac{\rho}{m}$$

Fator de utilização do servidor pode ser considerado como a fração esperada de servidores ocupados quando cada servidor possui a mesma distribuição de tempo de serviço.

Fator de utilização do servidor pode também ser considerado como a fração esperada da capacidade do sistema que está em uso.

Diagrama de estado para a fila M/M/*m*:



Probabilidade de transição de estado:

Quando $n \leq m$, para a condição de equilíbrio temos:

$$\lambda p_{n-1} + (n+1)\mu p_{n+1} = (\lambda + n\mu) p_n \qquad \text{para } n \geq 1$$

ou usando $\rho = \lambda/\mu$, temos:

$$\rho p_{n-1} + (n+1) p_{n+1} = (\rho + n) p_n \qquad \text{para } n \geq 1$$

Repetindo o mesmo processo recursivo utilizado para fila M/M/1, obtemos:

$$p_n = \frac{\rho^n}{n!} p_0 \qquad \text{para } n \le m$$

De modo similar, para $n \ge m$, temos a equação de equilíbrio detalhada para o caso de equilíbrio:

$$\lambda p_{n-1} + m\mu p_{n+1} = \left(\lambda + m\mu\right) p_n$$

ou usando $\rho = \lambda/\mu$, temos:

$$\rho p_{n-1} + m p_{n+1} = \left(\rho + m\right) p_n \qquad \text{para } n \ge m$$

Resolvendo recursivamente, obtemos:

$$p_n = \frac{\rho^n}{m!\ m^{n-m}} p_0 \qquad \text{para } n \ge m$$

Sabendo que

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

podemos obter a expressão de p_0 utilizando as duas expressões de p_n:

$$p_0 = \left[\sum_{n=0}^{m-1} \frac{\rho^n}{n!} + \sum_{n=m}^{\infty} \frac{\rho^n}{m! m^{n-m}} \right]^{-1} = \left[\sum_{n=0}^{m-1} \frac{\rho^n}{n!} + \frac{\rho^m}{m!} \frac{1}{\left(1-\rho/m\right)} \right]^{-1}$$

Relação utilizada para $x < 1$:

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

Relação utilizada para $q \neq 1$:

$$\sum_{k=1}^{S} a q^{k-1} = \frac{a\left(q^s - 1\right)}{q-1}$$

Valores de medidas de desempenho do sistema, N_q, T_q, N e T:

O comprimento da fila é obtido considerando os estados para $n \geq m$, pois uma fila se formará somente se todos os m servidores estiverem ocupados. Por definição:

$$N_q = E\left[\tilde{N}_q\right] = \sum_{n=m}^{\infty}(n-m)\,p_n = \sum_{k=0}^{\infty} k p_{m+k} = \sum_{k=0}^{\infty} k \frac{\rho^{m+k}}{m!\,m^k} p_0 = \frac{\rho^m\left(\rho/m\right)}{m!\left(1-\rho/m\right)^2} p_0$$

$$\boxed{N_q = \frac{\rho^m\left(\rho/m\right)}{m!\left(1-\rho/m\right)^2} p_0}$$

Utilizando a fórmula de Little:

$$T_q = \frac{N_q}{\lambda} \qquad N = T\lambda$$

onde, sabendo que o tempo médio de processamento $E[s] = 1/\mu$, temos:

$$T = T_q + T_s = T_q + E\left[s\right] = T_q + \frac{1}{\mu}$$

Probabilidade de todos os m servidores estarem ocupados tal que a mensagem que chega deve entrar na fila:

$$P\left[\text{entrar na fila}\right] = \sum_{n=m}^{\infty} p_n = \sum_{n=m}^{\infty} p_0 \frac{\rho^n}{m!\, m^{n-m}} = \frac{\dfrac{\rho^m}{m!}\dfrac{1}{\left(1-\rho/m\right)}}{\displaystyle\sum_{n=0}^{m-1}\frac{\rho^n}{n!} + \frac{\rho^m}{m!}\frac{1}{\left(1-\rho/m\right)}}$$

Fórmula *C* de Erlang ou fórmula de atraso de Erlang:

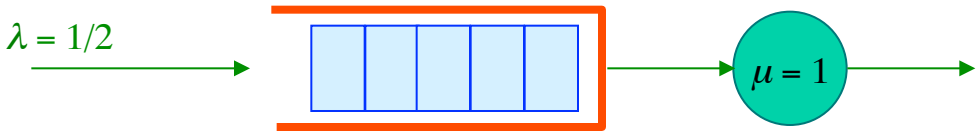
$$P\left[\text{entrar na fila}\right] = \frac{\dfrac{\rho^m}{m!}\dfrac{1}{\left(1-\rho/m\right)}}{\displaystyle\sum_{n=0}^{m-1}\frac{\rho^n}{n!} + \frac{\rho^m}{m!}\frac{1}{\left(1-\rho/m\right)}}$$

Probabilidade de nenhum servidor estar disponível para uma mensagem chegando a um sistema com m servidores.

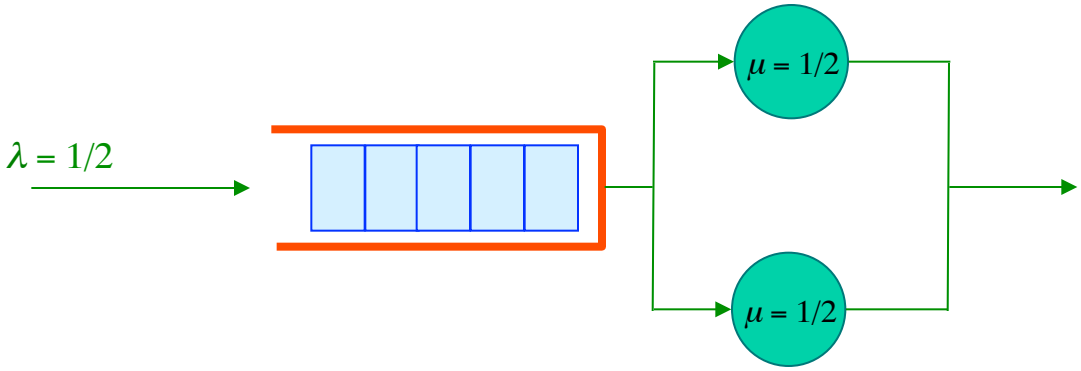
Exemplo:

Comparação do desempenho de atraso médio e de tempo de espera dos dois sistemas abaixo:

M/M/1:



M/M/2:



Para o M/M/1:

$$\rho = \frac{\lambda}{\mu} = \frac{1/2}{1} = 0{,}5$$

Tempo médio de espera:

$$T_q = \frac{\rho/\mu}{1-\rho} = \frac{0{,}5/1}{1-0{,}5} = 1 \text{ [s]}$$

Atraso médio total:

$$T = \frac{1/\mu}{1-\rho} = \frac{1/1}{1-1/2} = 2 \text{ [s]}$$

Para o M/M/2:

$$\rho = \frac{\lambda}{\mu} = \frac{1/2}{1/2} = 1$$

Probabilidade do sistema estar vazio:

$$p_0 = \left[\sum_{n=0}^{m-1}\frac{\rho^n}{n!} + \frac{\rho^m}{m!}\frac{1}{\left(1-\rho/m\right)}\right]^{-1} = \left[\sum_{n=0}^{1}\frac{1^n}{n!} + \frac{1^2}{2!}\frac{1}{\left(1-1/2\right)}\right]^{-1} = \left[\frac{1^0}{0!} + \frac{1^1}{1!} + \frac{1^2}{2!}2\right]^{-1} = \frac{1}{3}$$

Fórmula C de Erlang:

$$P\left[\text{entrar na fila}\right] = \frac{\dfrac{\rho^m}{m!}\dfrac{1}{\left(1-\rho/m\right)}}{\displaystyle\sum_{n=0}^{m-1}\frac{\rho^n}{n!}+\frac{\rho^m}{m!}\frac{1}{\left(1-\rho/m\right)}} = \frac{\dfrac{1^2}{2!}\dfrac{1}{\left(1-1/2\right)}}{\displaystyle\sum_{n=0}^{m-1}\frac{1^n}{n!}+\frac{1^2}{2!}\frac{1}{\left(1-1/2\right)}} = \frac{1}{3}$$

Tempo médio de espera:

$$T_q = \frac{\rho^m\left(\rho/m\right)}{m!\left(1-\rho/m\right)^2\lambda}p_0 = \frac{1^2\left(1/2\right)}{2!\left(1-1/2\right)^2 0{,}5}\cdot\frac{1}{3} = \frac{2}{3}\ \text{[s]}$$

Atraso médio total:

$$T = T_q + \frac{1}{\mu} = \frac{2}{3}+2 = \frac{8}{3}\ \text{[s]}$$

Conclusão:

Sistema M/M/1 possui menor atraso médio total e maior tempo médio de espera que o sistema M/M/2.

O aumento do número de servidores diminui o tempo de espera e aumenta o atraso total.

III.7 FILAS M/G/1

Envolve uma distribuição de tempo de serviço.

Assumimos que:

- Sistema caracterizado por um processo de chegada de Poisson com uma taxa média λ chegadas por segundo.

- Somente transições entre estados adjacentes são permitidas.

- As mensagens são processadas na base FCFS.

Entretanto, a distribuição de tempo de serviço possui uma forma arbitrária ou geral $B(t)$ com tempo médio de serviço $1/\mu$.

Os resultados mais utilizados para filas M/G/1 são as fórmulas de valor médio de Pollaczek-Khinchin (*P-K*) para o número médio de mensagens em uma fila e o atraso de tempo médio.

Razão da variância σ_b^2 do tempo de serviço pelo tempo médio de serviço ao quadrado:

$$C_b^2 = \frac{\sigma_b^2}{1/\mu^2} = \sigma_b^2 \mu^2$$

Número médio de mensagens no sistema é dado pela fórmula de valor médio *P-K*:

$$N = E\left[\tilde{N}\right] = \rho + \rho^2 \frac{1 + C_b^2}{2(1-\rho)}$$

Número médio de mensagens na fila:

$$N_q = N - N_s = N - \rho = \rho^2 \frac{1 + C_b^2}{2(1-\rho)}$$

$$N_s = \frac{\lambda}{\mu} = \rho$$

Exemplo: Comprimento das mensagens são exponencialmente distribuídas.

Então,

$$\sigma_b^2 = \frac{1}{\mu^2} \quad \Longrightarrow \quad C_b^2 = \sigma_b^2 \mu^2 = 1$$

Substituindo na equação de N, obtemos:

$$N = E\left[\tilde{N}\right] = \rho + \rho^2 \frac{1 + C_b^2}{2(1-\rho)} = \frac{\rho}{1-\rho}$$

que é idêntico ao número médio de mensagens em um sistema M/M/1, como esperado.

Exemplo: Comprimento das mensagens são fixas = tempo de serviço fixo. Então,

$$\sigma_b^2 = 0 \quad \Longrightarrow \quad C_b^2 = \sigma_b^2 \mu^2 = 0$$

Substituindo na equação de *N*, obtemos:

$$N = E\left[\tilde{N}\right] = \rho + \rho^2 \frac{1+0}{2(1-\rho)} = \frac{\rho}{1-\rho} - \frac{\rho^2}{2(1-\rho)}$$

Que indica que o número médio de mensagens é $\rho^2/[2(1-\rho)]$ menor que em um sistema M/M/1.

Este sistema é conhecido como um sistema de fila M/D/1.

Note que o nº médio de mensagens aumenta com o aumento de σ_b^2 .

O atraso médio de uma mensagem passando pelo sistema pode ser encontrado utilizando a fórmula de Little:

$$T = \frac{1}{\lambda} E\left[\tilde{N}\right] = \frac{1}{2\mu(1-\rho)}\left[2 - \rho\left(1 - \mu^2\sigma_b^2\right)\right]$$

Para comprimento de mensagens exponencial temos $\sigma_b^2\mu^2 = 1$, portanto:

$$T = \frac{1}{\mu(1-\rho)}$$

Para comprimento de mensagens fixo temos o mesmo atraso da fila M/M/1:

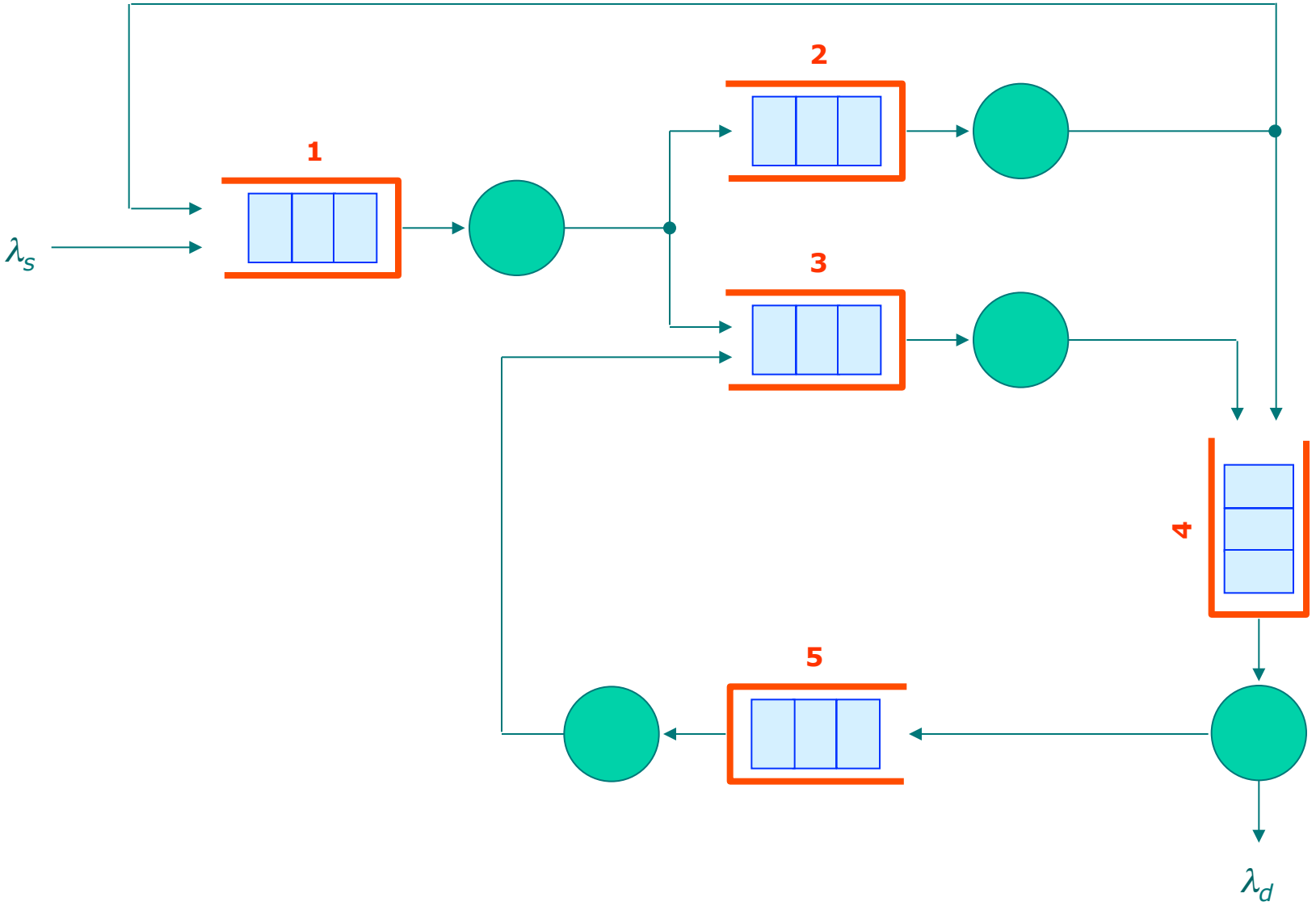
$$T = \frac{1}{\mu(1-\rho)}$$

III.8 REDES DE FILAS

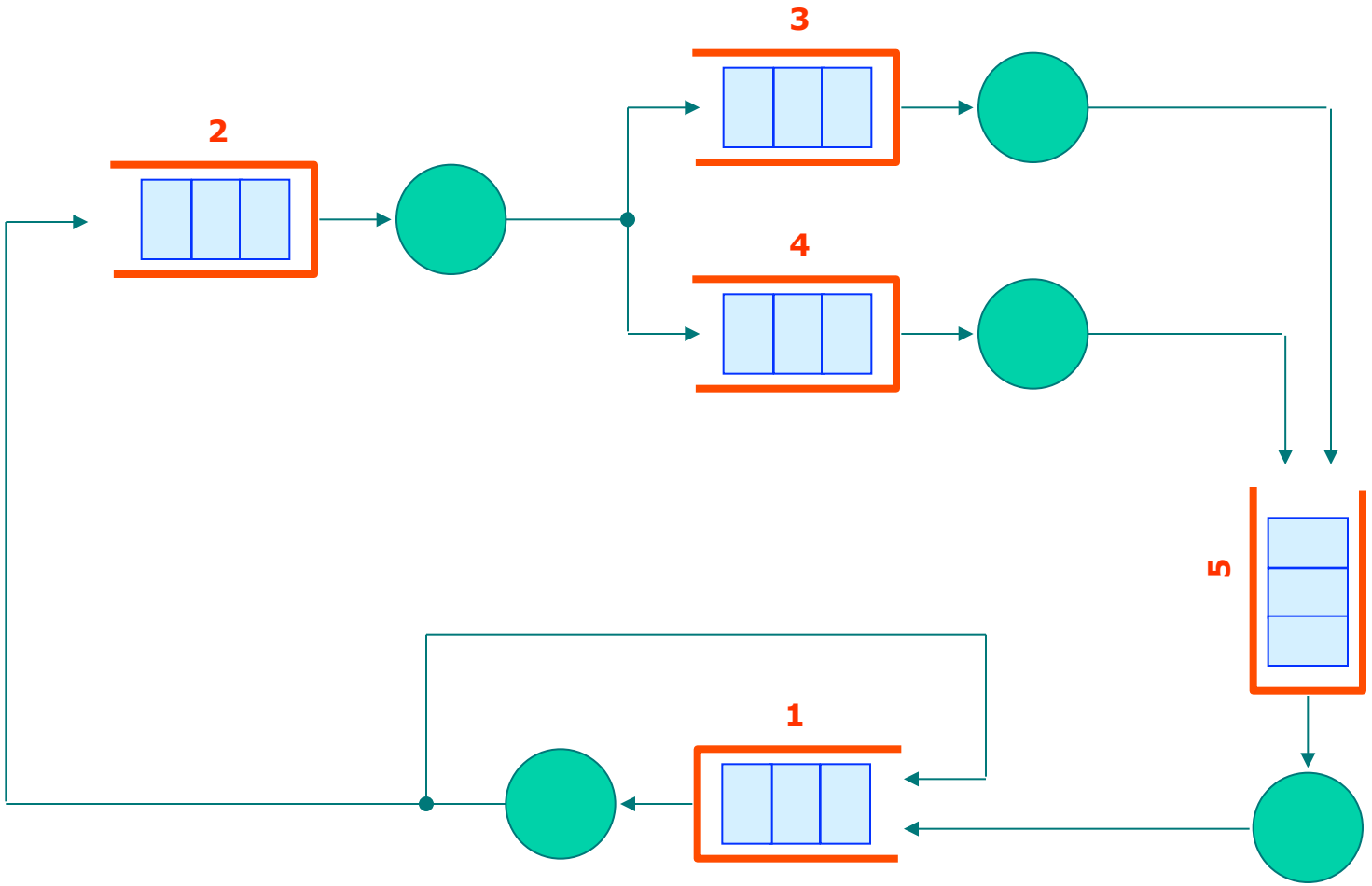
Classes de redes de fila:

- Sistemas de filas abertos, caracterizados pelo fato do número de mensagens no sistema não ser fixo, o tráfego pode entrar e sair do sistema.
- Sistemas de filas fechados, caracterizados pelo fato do número de mensagens no sistema ser fixo, não há chegadas nem partidas externas.

Rede de filas aberta:



Rede de filas fechada:



III.8.1 Restrições de Modelagem

Dificuldade no estabelecimento de um modelo analítico:

- Os tempos entre chegadas se tornam fortemente correlacionados com os tamanhos das mensagens.

Para transpor esta dificuldade consideraremos cada estágio da rede como uma fila M/M/1, baseado na seguinte assertiva:

Assertiva de independência de Kleinrock:

> Se uma nova escolha de tamanho de mensagem é feita independentemente de cada fila (isto é, de cada enlace de saída), então modelos de filas M/M/1 separados podem ser usados para cada enlace de comunicação entre nós.

Boa aproximação para sistemas com:

- chegadas de mensagens de Poisson nos pontos de entradas,
- tamanhos de mensagem exponencialmente distribuídos,
- uma rede densamente conectada,
- carga de tráfego de moderada a pesada.

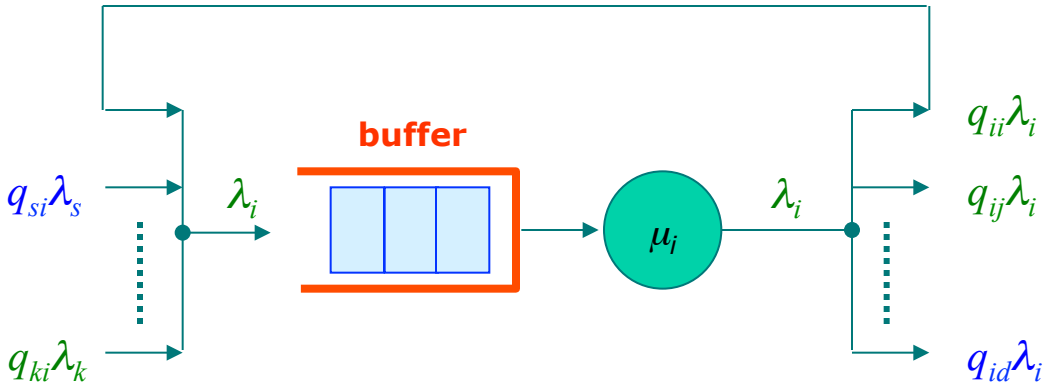
A modelagem de uma sequência de enlaces como filas M/M/1 independentes dá origem ao teorema de Jackson.

III.8.2 Teorema de Jackson

Este teorema diz que uma rede com serviço FCFS possui soluções que estão em uma forma de produto.

Consideramos um rede aberta com M filas.

Fila i típica:



q_{xy} = probabilidades de roteamento de que uma vez completado o serviço sobre a mensagem na fila x, esta é roteada para a fila y.

s = fonte

d = destino

λ_j = taxas de chegada com distribuição de Poisson.

μ_i = taxa de serviço da fila.

Seja o vetor de estado global $\mathbf{n} = (n_1, n_2, ..., n_M)$, onde n_i = estado da fila i.

Requisito de uma fila é a continuidade de fluxo na entrada e na saída.

Seja λ_i a taxa total de chegada na fila i da fonte e de todas as filas, então:

$$\lambda_i = q_{si}\lambda_s + \sum_{k=1}^{M} q_{ki}\lambda_k$$

Teorema de Jackson diz que a probabilidade de equilíbrio $p(\mathbf{n})$ de que a rede está no estado $\mathbf{n}$ é dada por:

$$p(\mathbf{n}) = p(n_1, n_2, \ldots, n_M) = p_1(n_1)\, p_2(n_2) \ldots p_M(n_M)$$

$p_i(n_i)$ = probabilidade de equilíbrio de que a fila i está no estado n_i.

Para provar o teorema de Jackson temos que verificar que a equação de equilíbrio global é satisfeita por uma solução na forma de produto como dada por:

$$p(\mathbf{n}) = p(n_1, n_2, \ldots, n_M) = p_1(n_1)p_2(n_2)\ldots p_M(n_M)$$

Igualando a taxa total de partida do estado **n** à taxa de entrada em **n**, temos a equação de balanço global:

$$\left[\lambda_s + \sum_{i=1}^{M} \mu_i\right] p(\mathbf{n}) = \lambda_s \sum_{i=1}^{M} q_{si}\, p(n_1, n_2, \ldots, n_i - 1, \ldots, n_M)$$

$$+ \sum_{i=1}^{M} q_{id}\mu_i\, p(n_1, n_2, \ldots, n_i + 1, \ldots, n_M)$$

$$+ \sum_{i=1}^{M}\sum_{j=1}^{M} q_{ij}\mu_j\, p(n_1, n_2, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_M)$$

Lado esquerdo da equação:

$$\left[\lambda_s + \sum_{i=1}^{M} \mu_i\right] p(\mathbf{n}) =$$ a taxa total de partida do estado **n**, desde que possa haver uma chegada com taxa λ ou uma partida a uma taxa μ_i de qualquer uma das M filas.

Lado direito da equação é a taxa total de chegada no estado **n**:

1º termo: refere-se às chegadas de uma fonte externa na fila i que está no estado $n_i - 1$. As taxas de chegada são $\lambda_s q_{si}$.

$$\lambda_s \sum_{i=1}^{M} q_{si} p(n_1, n_2, \ldots, n_i - 1, \ldots, n_M)$$

2º termo: é obtido das partidas da fila i diretamente para o destino d. As taxas de partida da fila i são $\mu_i q_{id}$.

$$\sum_{i=1}^{M} q_{id}\mu_i p\left(n_1, n_2, \ldots, n_i + 1, \ldots, n_M\right)$$

3º termo: descreve as transições da fila j, que originalmente está no estado $n_j + 1$, para a fila i, que está no estado $n_i - 1$.

$$\sum_{i=1}^{M}\sum_{j=1}^{M} q_{ij}\mu_j p\left(n_1, n_2, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_M\right)$$

Para resolver a equação anterior usamos

$$p(\mathbf{n}) = p(n_1, n_2, \ldots, n_M) = p_1(n_1)p_2(n_2)\ldots p_M(n_M)$$

para eliminar q_{si}.

Na expressão resultante, substituímos a relações:

$$\lambda_i p(n_1, n_2, \ldots, n_i - 1, \ldots, n_M) = \mu_i p(n_1, n_2, \ldots, n_i, \ldots, n_M)$$

$$\lambda_j p(n_1, n_2, \ldots, n_i - 1, \ldots, n_M) = \mu_j p(n_1, n_2, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_M)$$

$$\lambda_i p(n_1, n_2, \ldots, n_i, \ldots, n_M) = \mu_i p(n_1, n_2, \ldots, n_i + 1, \ldots, n_M)$$

Conceito da reversibilidade de Reich:

Em equilíbrio, transições de um estado para outro em tempo reverso ocorrem com as mesmas taxas que para as mesmas transições em tempo direto.

Então, a sequência de chegada original é equivalente em todos os aspectos a sequência de partida em tempo reverso, o que é a essência da equação:

$$\lambda_i p(n_1, n_2, \ldots, n_i - 1, \ldots, n_M) = \mu_i p(n_1, n_2, \ldots, n_i, \ldots, n_M)$$

Realizando as substituições, a equação

$$* \qquad \left[\lambda_s + \sum_{i=1}^{M} \mu_i\right] p(\mathbf{n}) = \lambda_s \sum_{i=1}^{M} q_{si}\, p(n_1, n_2, \ldots, n_i - 1, \ldots, n_M)$$

$$+ \sum_{i=1}^{M} q_{id} \mu_i\, p(n_1, n_2, \ldots, n_i + 1, \ldots, n_M)$$

$$+ \sum_{i=1}^{M} \sum_{j=1}^{M} q_{ij} \mu_j\, p(n_1, n_2, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_M)$$

se reduz à condição de conservação de fluxo da fonte para o destino:

$$\lambda_s = \sum_{i=1}^{M} q_{id} \lambda_i$$

Então, a condição de equilíbrio em * é satisfeita por:

$$\lambda_i\, p(n_1, n_2, \ldots, n_i - 1, \ldots, n_M) = \mu_i\, p(n_1, n_2, \ldots, n_i, \ldots, n_M)$$

Para a fila i temos então,

$$p(\mathbf{n}) = \left(\frac{\lambda_i}{\mu_i}\right) p\left(n_1, n_2, \ldots, n_i - 1, \ldots, n_M\right)$$

Repetindo este processo n_i vezes, obtemos:

$$p(\mathbf{n}) = \left(\frac{\lambda_i}{\mu_i}\right)^{n_i} p\left(n_1, n_2, \ldots, n_i = 0, \ldots, n_M\right)$$

Seguindo estes mesmos passos para todas as outras filas, temos:

$$p(\mathbf{n}) = \prod_{i=1}^{M} \left(\frac{\lambda_i}{\mu_i}\right)^{n_i} p(\mathbf{0})$$

$p(\mathbf{0})$ = probabilidade de todas as M filas estarem vazias.

Para encontrar $p(\mathbf{0})$, fazemos $\rho_i = \lambda_i/\mu_i$ e somamos a equação anterior sobre todos os estados possíveis, igualando o resultado a 1:

$$\sum_{\mathbf{n}} p(\mathbf{n}) = p(\mathbf{0}) \sum_{\mathbf{n}} \left[\prod_{i=1}^{M} \rho_i^{n_i} \right] = 1$$

O termo entre colchetes deve ser finito para haver solução, logo:

$$\sum_{\mathbf{n}} \left[\prod_{i=1}^{M} \rho_i^{n_i} \right] = \prod_{i=1}^{M} \sum_{n_i=0}^{\infty} \rho_i^{n_i} = \prod_{i=1}^{M} \left(\frac{1}{1-\rho_i} \right)$$

Assim, chegamos ao teorema de Jackson:

$$p(\mathbf{n}) = \prod_{i=1}^{M} \rho_i^{n_i} (1-\rho_i) = \prod_{i=1}^{M} p_i(n_i)$$

O teorema de Jackson pode ser estendido para redes M/M/m onde cada fila i possui m_i servidores, ou seja

$$p(\mathbf{n}) = p(n_1, n_2, \ldots, n_M) = p_1(n_1)p_2(n_2)\ldots p_M(n_M)$$

onde a equação usada para os $p_i(n_i)$ neste caso é dada por:

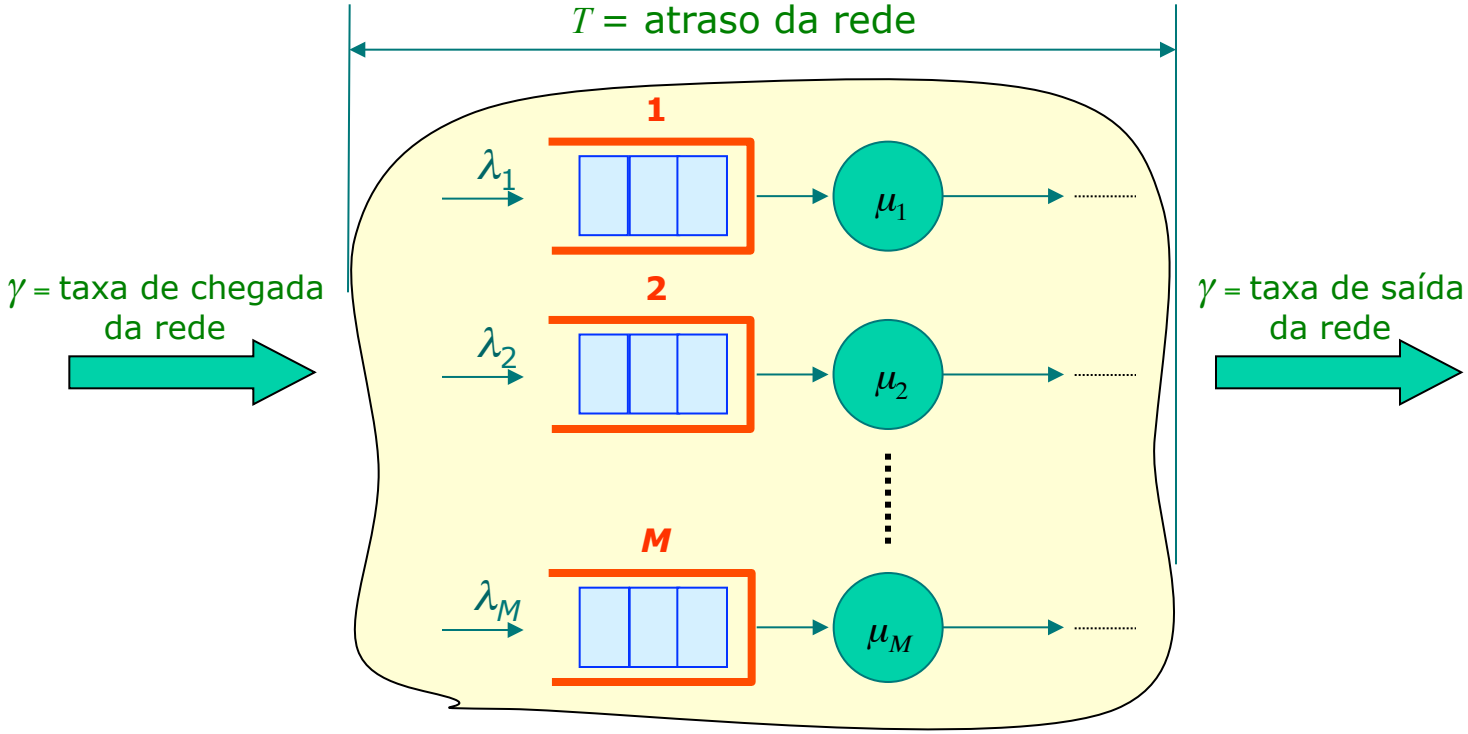
$$p_i(n_i) = \frac{\rho^{n_i}}{m_i! m_i^{m_i - n_i}} p_0 \qquad \text{para } n_i \geq m_i$$

III.8.3 Aplicação em uma Redes de Filas

Aplicação da equação:

$$p(\mathbf{n}) = p(n_1, n_2, \ldots, n_M) = p_1(n_1)p_2(n_2)\ldots p_M(n_M)$$

Consideramos o atraso médio de uma rede ampla com a média feita sobre todos os M enlaces.



Rede aberta em equilíbrio = coleção de M filas

γ_s = taxa de chegada (mensagens/segundo) associada ao percurso s.

Taxa chegada total na rede:

$$\gamma = \sum_s \gamma_s$$

Da fórmula de Little temos o atraso T médio da rede onde a média é feita para toda a rede:

$$\gamma T = E[\mathbf{n}]$$

$E[\mathbf{n}]$ = nº médio de mensagens enfileiradas ou em serviço na rede:

$$E[\mathbf{n}] = \sum_{i=1}^{M} E[n_i]$$

$E[n_i]$ = nº médio de mensagens enfileiradas ou em serviço no nó i.

Para uma fila M/M/1, temos:

$$E\left[n_i\right] = \lambda_i T_i = \frac{\lambda_i}{\mu_i - \lambda_i}$$

Então, o atraso médio é dado por:

$$T = \frac{1}{\gamma} \sum_{i=1}^{M} \frac{\lambda_i}{\mu_i - \lambda_i}$$

A equação acima não leva em consideração o atraso de propagação nos enlaces. Se este for incluído, então para um atraso de propagação T_{di} sobre o enlace i, obtemos:

$$T = \frac{1}{\gamma} \sum_{i=1}^{M} \left(\frac{\lambda_i}{\mu_i - \lambda_i} + \lambda_i T_{di} \right)$$