Redes de Filas e Leis Operacionais

1 Introdução

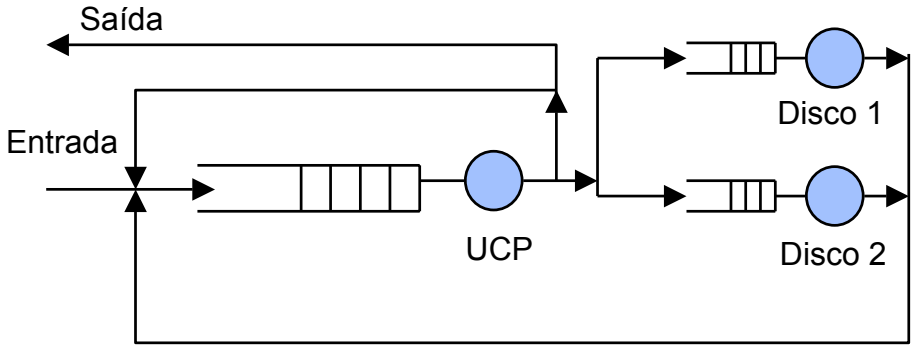
O objetivo é apresentar a solução de sistemas que envolvem múltiplas filas. O enfoque dado é o de aplicar as soluções encontradas em sistemas reais. Neste e nos próximos capítulos serão estudadas soluções exatas e aproximadas.

Os sistemas de filas são classificados em:

- Redes abertas;
- Redes fechadas;
- Redes mistas.

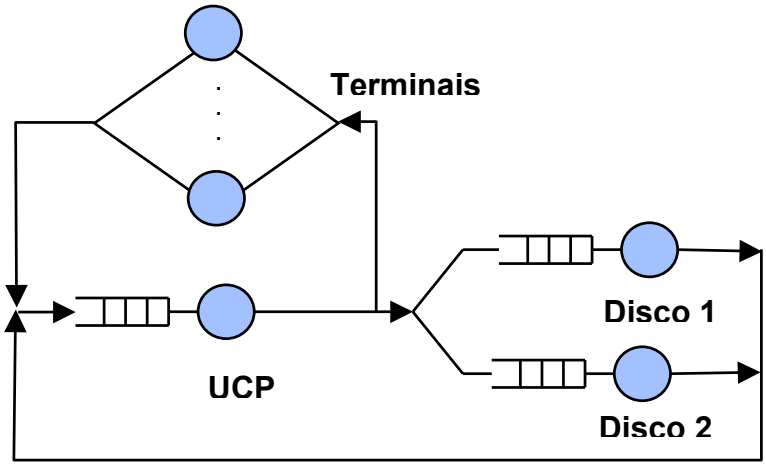
Exemplo 1: Rede Aberta de Filas

Um sistema computacional com um processador e dois discos para processamento em "Batch" pode ser representado da seguinte forma:



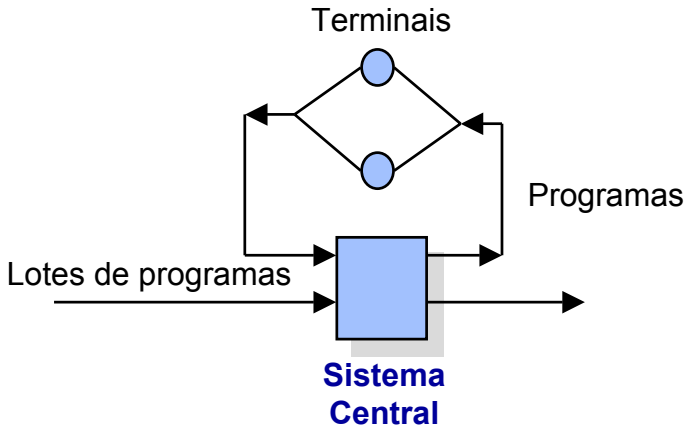
Exemplo 2: Rede Fechada de Filas

Um sistema computacional com um processador e dois discos para processamento em "Time-sharing" pode ser representado da seguinte forma:



Exemplo 3: Rede Mista de Filas

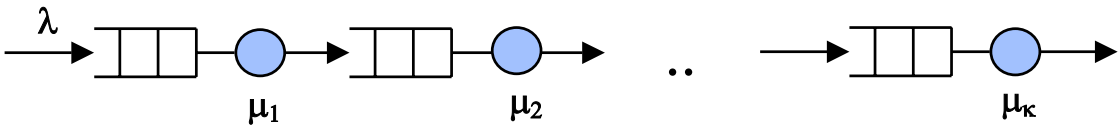
Um sistema computacional com um processador para processamento em "Batch"e em "Time-sharing" pode ser representado da seguinte forma:



2 Solução na Forma de Produto:

Exemplo 4: Associação série de filas

A forma mais simples de uma associação de filas é uma associação em série.



Supondo que as filas estão em equilíbrio, a taxa de chegada (λ) é igual a taxa de saída em todas as filas. Neste caso o fator de utilização da i-ésima fila é:

$$\rho_i = \lambda/\mu_i$$

A probabilidade de se ter n_i usuários na fila i é:

$$p_i(n_i) = (1 - \rho_i) * \rho_i^{ni}$$

A probabilidade conjunta de se ter n_i usuários na fila i, para $i = 1,2,...,M$ é dada por:

$$P(n_1,n_2,...,n_M) = p_1(n_1)p_2(n_2)...p_M(n_M) = \prod_{i=1}^{M}(1-\rho_i)\rho_i^{n_i}$$

Esta fórmula é uma solução na forma de produto que, no caso geral, possui a seguinte expressão:

$$P(n_1,n_2,...,n_M) = \frac{1}{G(N)} * \prod_{i=1}^{M} f_i(n_i)$$

onde G(N) é uma constante de normalização função do número total de usuários no sistema N.

2.1 Condições para existência de Forma de Produto

Diversos autores verificaram condições para que a rede de filas apresente solução na forma de produto. Os principais resultados são os seguintes:

Condição de Jackson

Existe solução na forma de produto para qualquer rede aberta com filas possuindo m servidores exponenciais.

Condição de Gordon e Newell

Existe para qualquer rede fechada com filas possuindo m servidores exponenciais.

Condições Baskett, Chandy, Muntz e Palácios (BCMP)

Existe solução em forma de produto qualquer rede aberta ou fechada com filas obedecendo às seguintes restrições:

a) **Disciplina de serviço**: Todas as filas possuem uma das seguintes disciplinas: FCFS ou PS ou IS ou LCFS-PR.
b) **Classes de Usuários**: Usuários não mudam de classe enquanto estão esperando ou sendo atendidos. Mudam de classe somente quando terminam o seu serviço.
c) **Distribuição de tempo de serviço**: Para filas FCFS, todos os servidores precisam ser idênticos e exponencialmente distribuídos para todas as classes. Para outros centros de serviços onde as distribuições de probabilidade possuem transformada de Laplace racional, diferentes classes podem ter diferentes distribuições.
d) **Serviço dependente do estado**: Para disciplinas FCFS o tempo de serviço só pode depender do número total de usuários no sistema. Para disciplinas PS, ou LCFS-PR, ou IS o tempo de serviço pode depender também do número de usuários de sua classe, mas não do número das outras classes.
e) **Processo de chegada**: Nas redes abertas o processo de chegada deve ser exponencialmente distribuído. Sistemas com chegadas em lotes não são válidos. A taxa de chegada pode ser dependente do estado do sistema.

A rede pode ser aberta em relação a algumas classes e fechada em relação a outras classes.

Condições de Denning e Buzen

Redes de fila Não-Markovianas possuem solução em forma de produto quando verificam as seguintes condições:

a) **Fluxo de usuários balanceado**: Para cada classe, o número de chegadas é igual ao número de partidas.

b) **Eventos únicos**: Não podem existir eventos múltiplos simultâneos.
c) **Homogeneidade de dispositivos**: A taxa de serviço para uma particular classe não depende do estado do sistema a não ser do número total de usuários e do número de usuários de sua classe no sistema.

3 Leis Operacionais

As leis operacionais dos sistemas de filas são relações que existem entre as grandezas diretamente mensuráveis destes sistemas. Algumas das grandezas que podem ser diretamente mensuráveis nos sistemas de fila são:

A_i : Número de chegadas;
C_i : Número de partidas;
B_i : Tempo ocupado.

Valores derivados destas grandezas mensuráveis:

λ_i : Taxa de chegada = A_i/T
X_i : Vazão = C_i/T
U_i : Fator de Utilização = B_i/T
S_i : Tempo médio de serviço = B_i/C_i

Observe que estas grandezas podem assumir diferentes valores em diferentes períodos de observação. Porém, existem certas relações que permanecem válidas para cada período de observação. Estas relações são as chamadas **Leis Operacionais** dos Sistemas de Filas.

Lei da Utilização

Dado um número de partidas C_i, um tempo de ocupação B_i, de um sistema de filas i durante um intervalo de observação **T**, a seguinte relação é válida:

$$U_i = (B_i/T) = (C_i/T)*(B_i/C_i) \text{ ou}$$

$$U_i = X_i*S_i$$

Lei do Fluxo

Esta lei correlaciona a vazão global do sistema com as vazões de seus subsistemas. Em uma rede aberta de filas, o número de usuários partindo da rede na unidade de tempo define a sua vazão. Numa rede fechada, a taxa com que se cicla no sistema define a sua vazão.

Se num dado período T de observação, o número de usuários que entraram é igual ao número de usuários que saíram do sistema, isto é:

$$A_i = C_i$$

pode-se dizer que este sistema satisfaz a hipótese de fluxo balanceado.

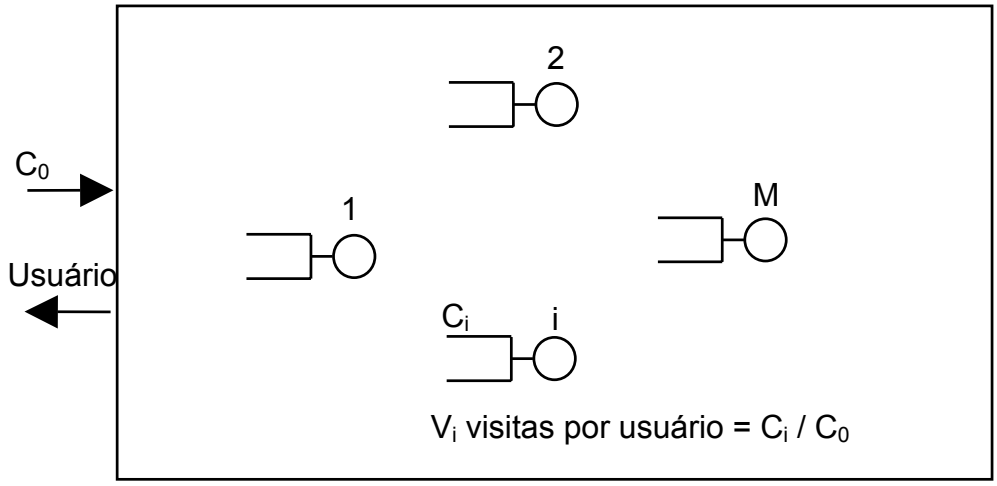
Se o intervalo de observação é grande, C_i tende a se aproximar de A_i.

Suponha que cada usuário faça V_i visitas ao i-ésimo subsistema. Se o fluxo deste sistema é balanceado, o número de usuários C_0 que entram ou saem, e o número de visitas ao i-ésimo subsistema estão relacionados pela seguinte expressão:

$$C_i = C_0 * V_i \qquad \text{ou} \qquad V_i = C_i / C_0$$

A variável V_i representa a taxa de visitas ao sub-sistema i para cada usuário.

Sistema de Filas



A vazão global do sistema durante este período de observação é dada por:

X : Vazão do Sistema = C_0/T

A vazão do i-ésimo subsistema é dada por:

X_i : Vazão do sub-sistema i = $C_i/T = (C_i/C_0)*(C_0/T)$

Isto é,

$$X_i = X * V_i \qquad \text{Lei de Fluxo.}$$

Combinando a lei de utilização com a lei de Fluxo tem-se:

$$U_i = X_i * S_i = X * V_i * S_i$$

ou

$$U_i = X * D_i$$

onde $D_i = V_i * S_i$ é chamado de demanda total sobre o i-ésimo sub-sistema. O subsistema que possuir o maior D_i será o gargalo do sistema.
A taxa de visitas é uma das maneiras de se especificar o roteamento dos usuários numa rede de filas. Uma outra forma é se especificar as probabilidades de transição p_{ij} de um usuário ao terminar o serviço em i se mover para j.

Num sistema com o fluxo balanceado tem-se:

$$C_j = \sum_{i=0}^{M} C_i p_{ij}$$

p_{i0} é a probabilidade do usuário deixar o sistema tendo terminado o serviço em i;
C_0 representa o número de usuários que entraram ou sairam do sistema;

Dividindo ambos os lados da relação por C_0 tem-se:

$$V_j = \sum_{i=0}^{M} V_i p_{ij}$$

Como a tarefa de um usuário termina ao sair do sistema, então

$$V_0 = 1.$$

As duas equações anteriores permitem que se obtenha as relações entre V_i e p_{ij}.

A lei de Little já foi vista e é expressa por:

$$Q_i = \lambda_i * R_i$$

onde Q_i é o número de usuários em i e Ri é o tempo gasto em i;

Para o caso de sistemas com fluxo balanceado pode-se escrever:

$$Q_i = X_i * R_i$$

onde X_i é a vazão em i.

Lei do Tempo de Resposta

Todo sistema de "Time-Sharing" pode ser dividido em dois subsistemas: os Terminais e o Sistema Central.

A lei de Little pode ser aplicada para qualquer destes subsistemas desde que ele possua fluxo balanceado:

$$Q = X * R \qquad \text{para o sistema Central.}$$

Conhecendo-se o número de usuários em cada um dos subsistemas do Sistema Central, pode-se escrever:

$$Q = Q_1 + Q_2 + \ldots + Q_M$$

como

$$Q_i = X_i * R_i$$

$$XR = X_1R_1 + X_2R_2 + \ldots + X_MR_M$$

dividindo-se ambos os lados por X e usando a lei do fluxo, tem-se:

$$R = \sum_{i=1}^{M} R_i V_i$$ **Lei do Tempo de Resposta**

Lei do Tempo de Resposta Interativo

Num sistema interativo o tempo em que um usuário gasta pensando antes de fornecer uma nova requisição ao sistema é Z. Se o tempo de resposta do sistema é R, então o tempo de um ciclo completo pelo sistema é:

$$(R + Z)$$

Cada usuário produz T/(R+Z) requisições ao sistema num intervalo de tempo **T**.

Em uum sistema com N usuários a vazão do sistema será dada por:

$$X = \{N[T/(R+Z)]/T\} = N/(R+Z)$$

ou

$$R = (N/X) - Z$$ **Lei do Tempo de Resposta Interativo**.

Análise de Gargalo

Em um sistema o dispositivo gargalo é aquele que possui a maior demanda de serviço $\mathbf{D_i}$, ou equivalentemente, o maior fator de utilização U_i. Suponha que o elemento gargalo seja **b**. Isto implica em $D_b = D_{max}$, onde D_{max} é o maior valor entre $D_1, D_2, \ldots, D_M$.

Sendo $U_b = X D_{max} \leq 1$ tem-se que $X \leq 1/D_{max}$ que é a vazão máxima possível.

Quando N=1 tem-se que $R(1) = D = \sum_{i=1}^{M} D_i$ que é o menor atraso possível.

Assim, $R(N) \geq D$

Então $R(N) = N/X(N) - Z \geq N D_{max} - Z$

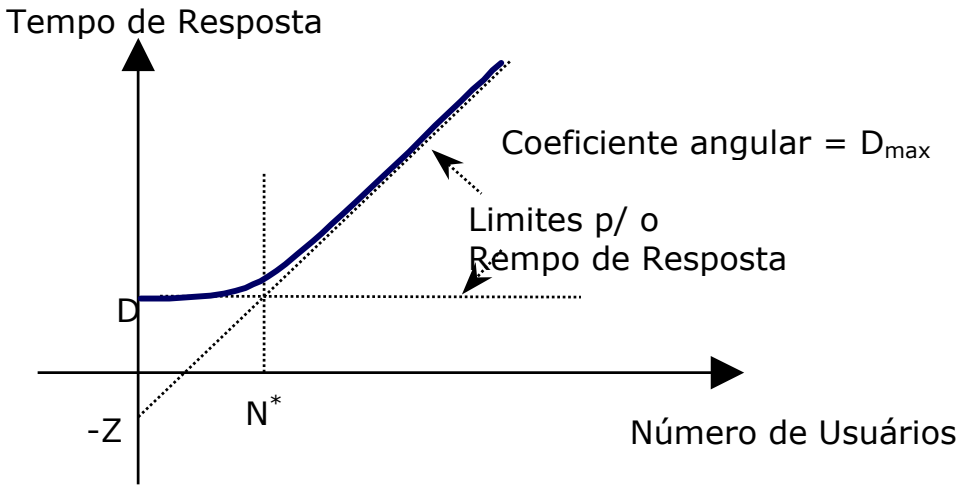
e $X(N) = N/[R(N) + Z] \leq N / [D + Z]$

Desta forma, a vazão e o tempo de resposta do sistema são limitados pelos seguintes valores:

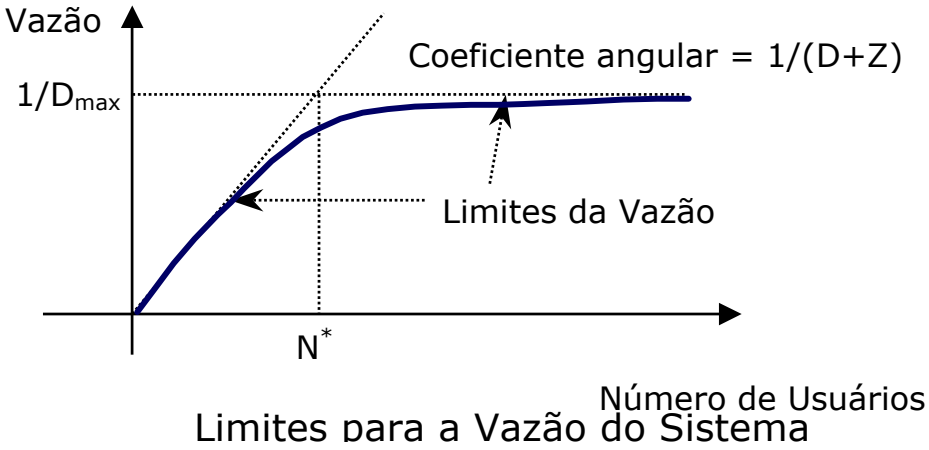
$$X(N) \leq \min \{ (1/D_{max}),(N/(D+Z)) \}$$

$$R(N) \geq \max\{ D,(ND_{max} - Z) \}$$

onde $D = \sum_{i=1}^{M} D_i$ é a soma da demanda de serviço de todos os sub-sistemas exceto os terminais. Estas inequações são chamadas de limites assintóticos.



Limites do Tempo de Resposta



Limites para a Vazão do Sistema

O ponto de interseção das duas retas limites é chamado **"joelho"** do sistema e é dado por:

$$N^* = (D+Z)/D_{max}$$

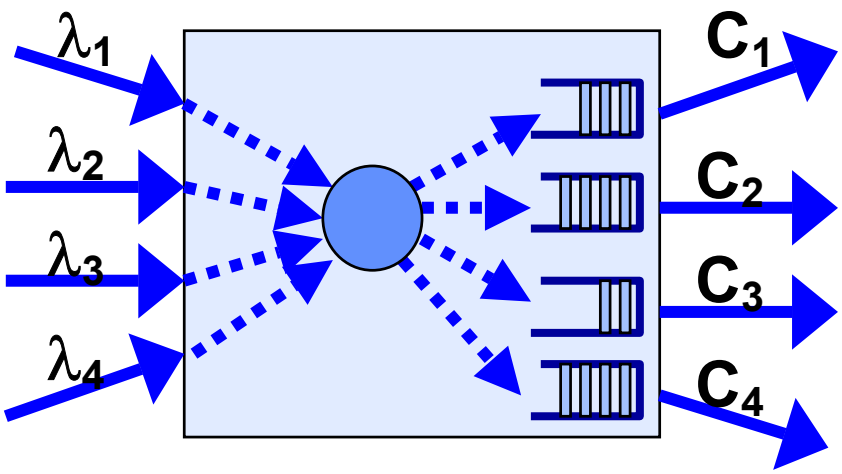
onde N^* é o número de usuários no "joelho" do sistema.

Se o número de usuários no sistema for maior que N*, pode-se dizer com certeza que existirá espera em algum lugar do sistema.

4 Estudo de Caso 1: Rede de Concentradores de Comunicação

4.1 Cálculo de Atrasos

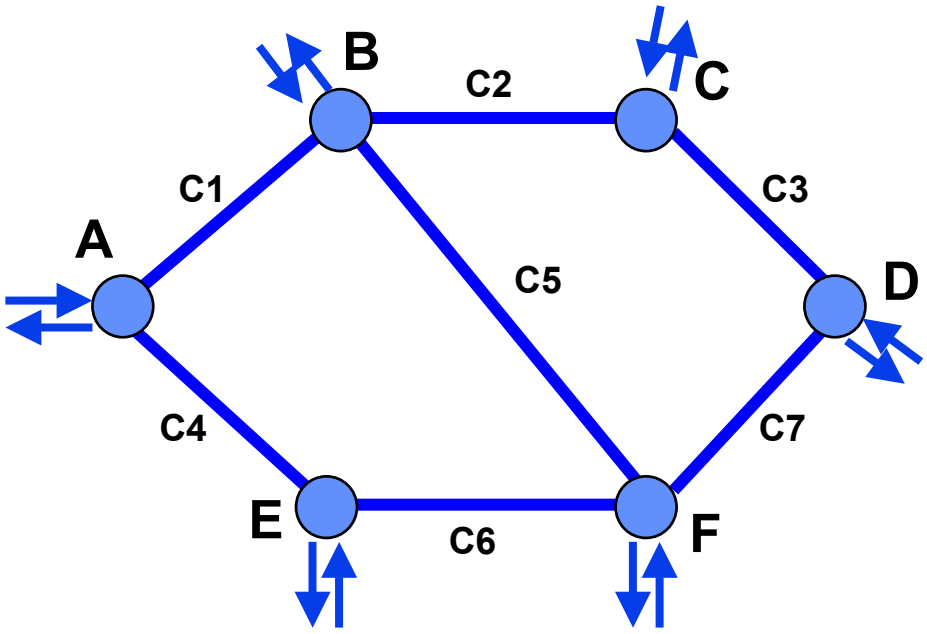
Um concentrador de comunicação, ou roteador pode ser representado de forma simplificada de acordo com o diagrama a seguir com uma fila para cada linha de saída:



C_i - capacidade da linha **i** de saída em Kbps

A rede de comunicação de dados é formada pela interconexão de concentradores através de linhas com capacidades expressas em Kbps (considerar K=1024bits). As linhas de comunicação são Full-duplex com as capacidades indicadas nas duas direções.

O tamanho do pacote tem distribuição exponencial com média 0,8 Kbits. Em cada nó entra tráfego externo destinado aos demais nós. A matriz de tráfego (γ_{ij}) define o tráfego que entra no nó i com destino ao nó j em pacotes/seg. Este tráfego tem distribuição exponencial com as médias indicadas pelos elementos da matriz.



Matriz de Tráfego

$(\gamma_{ij}) =$

Nó Origem \ Nós Destinos	A	B	C	D	E	F
A		8	5	2	6	3
B	8		7	4	3	5
C	5	7		4	2	3
D	2	4	4		2	5
E	6	3	2	2		4
F	3	5	3	5	4	

Tabelas de Roteamento

Nó Destino	Tab A	Tab B	Tab C	Tab D	Tab E	Tab F
A	-	R1	R2	R3	R4	R6
B	C1	-	R2	R3	C6	R5
C	C1	C2	-	R3	C6	R7
D	C1	C2	C3	-	C6	R7
E	C4	C5	C3	C7	-	R6
F	C4	C5	C3	C7	C6	

Nas tabelas de roteamento acima, indicamos Rk como sendo o canal Ck no sentido reverso.

Caminhos para os nós Destinos

Nó atual	Nó A	Nó B	Nó C	Nó D	Nó E	Nó F
A	-	C1	C1-C2	C1-C2-C3	C4	C4-C6
B	R1	-	C2	C2-C3	C5-R6	C5
C	R2-R1	R2	-	C3	C3-C7-R6	C3-C7
D	R3-R2-R1	R3-R2	R3	-	C7-R6	C7
E	R4	C6-R5	C6-R7-R3	C6-R7	-	C6
F	R6-R4	R5	R7-R3	R7	R6	

λ_k é o tráfego que passa pela linha k e é calculado como a soma dos γ_{ij} tais que k está no caminho entre i e j. Neste caso, os servidores são as linhas e a capacidade de serviço da linha é μC_k (pacotes/segundo) onde $1/\mu$=0,8 Kbits por pacote.

O atraso médio T_k da linha k é calculado pela fórmula do sistema M/M/1 como

$$T_k = \frac{1}{\mu C_k - \lambda_k}$$

k	Linha	λ_k (pacotes/seg)	C_k (kbps)	μC_k (pac/seg)	T_k (ms)
1	C1	8+5+2=15	20	25,0	100
2	C2	5+2+7+4=18	30	37,5	51
3	C3	2+4+4+2+3=15	20	25,0	100
4	C4	6+3=9	20	25,0	63
5	C5	3+5=8	10	12,5	222
6	C6	3+4+3+2+2=14	20	25,0	91
7	C7	2+3+2+5=12	20	25,0	77
8	R1	8+5+2=15	20	25,0	100
9	R2	5+2+7+4=18	30	37,5	51
10	R3	2+4+4+2+3=15	20	25,0	100
11	R4	6+3=9	20	25,0	63
12	R5	3+5=8	10	12,5	222
13	R6	3+4+3+2+2=14	20	25,0	91
14	R7	2+3+2+5=12	20	25,0	77
		$\Sigma\lambda_k$=182			

Chamamos de γ e λ

$$\gamma = \sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij} \qquad \lambda = \sum_{k=1}^{m}\lambda_k$$

O número médio de pulos (hops) por pacote é $\bar{n} = \lambda / \gamma$

λ é relacionado ao tráfego γ_{ij} por

$$\lambda = \sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\gamma_{ij}$$

onde h_{ij} é o número de pulos no caminho entre i e j.

O atraso médio total T é

$$T = \bar{n}\sum_{k=1}^{m}\frac{\lambda_k T_k}{\lambda} = \bar{n}\sum_{k=1}^{m}\frac{\lambda_k / \lambda}{\mu C_k - \lambda_k}$$

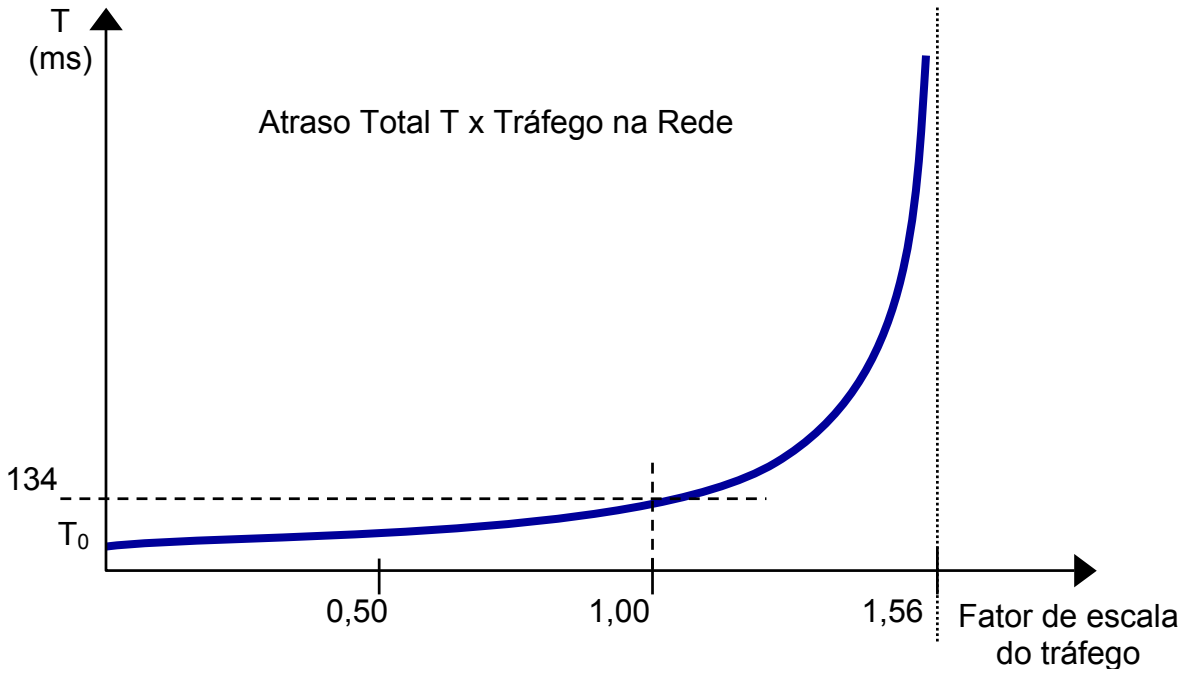
onde λ_k/λ é a proporção do tráfego total que passa pela linha k.

No exemplo $\lambda = 182$ e $\gamma = 126$ pacotes/seg,

$$\overline{n} = 182/126 = 1.44 \qquad \text{e} \qquad T=134 \text{ ms}$$

Deve ser observado que, se para algum k, $\mu C_k = \lambda_k$ então o atraso total T se torna infinito, isto é, basta o tráfego em um canal atingir o valor crítico para o tempo médio total se tornar infinito. No exemplo, nenhum λ_k causa esta condição. Contudo, se aumentarmos o tráfego na rede, iremos verificar que existe um canal que é mais vulnerável a variações do tráfego.

A linha que satura antes das demais é a que possui menor relação $\mu C_k / \lambda_k$. No exemplo, o canal que é mais vulnerável a variações do tráfego é o C5=BF e R5=FB onde $\mu C_k / \lambda_k$= 1,56. Se aumentarmos todos os valores da matriz de tráfego por este fator de escala atingiremos o valor de saturação da rede.



O gráfico mostra os valores de T com o tráfego variando de acordo com fatores de escala sendo que com o fator de escala 1,56 o atraso total se torna infinito.

Outro valor interessante de se observar é T_0 que é o atraso médio total quando não existe tráfego na rede. Para isto consideramos $\lambda_k = 0$.

$$T_0 = \overline{n} \sum_{k=1}^{m} \frac{\lambda_k / \lambda}{\mu C_k}$$

No exemplo, $T_0 = 38$ ms.

4.2 Designação de Capacidade

O cálculo de designação de capacidade permite responder à seguinte questão:

> **"Dada uma rede de concentradores, quais são as capacidades de canais necessárias para que o atraso se mantenha dentro de um limite e o custo não ultrapasse um determinado valor".**

Devemos considerar uma função de custo, tal como a seguinte função de custo linear:

$$(1) \qquad \text{Custo} = \sum_{i=1}^{m} (d_i C_i + x_i)$$

onde

d_i é o custo da linha i por bps
C_i é a capacidade da linha i a ser determinada
x_i é uma constante de custo associada à linha i não dependente da capacidade.

Queremos diminuir o custo sujeito à seguinte limitação do atraso total:

$$\frac{1}{\gamma} \sum_{i=1}^{m} \frac{\lambda_i}{\mu C_i - \lambda_i} = T$$

Neste caso T é uma constante que define o atraso máximo que será tolerado.

Este problema é resolvido por técnicas de otimização. No caso será utilizado o método de **Lagrange** definindo-se o multiplicador β e a função F a ser minimizada.

$$F = \sum_{i=1}^{m} (d_i C_i + x_i) + \beta \left[\frac{1}{\gamma} \sum_{i=1}^{m} \frac{\lambda_i}{\mu C_i - \lambda_i} - T \right]$$

Como a expressão entre colchetes é zero, multiplicá-la por β e somá-la à função custo não vai alterar a natureza da minimização. Esta será feita pelo cálculo de derivadas parciais.

$$\frac{\partial F}{\partial C_i} = d_i - \frac{\mu \beta \lambda_i}{\gamma} \left[\frac{1}{\mu C_i - \lambda_i} \right]^2 = 0$$

Esta equação pode ser re-escrita como:

$$(2) \qquad \frac{1}{\mu C_i - \lambda_i} = \sqrt{\gamma d_i / \mu \beta \lambda_i}$$

ou

$$(3) \qquad C_i = \frac{\lambda_i}{\mu} + \sqrt{\beta} \sqrt{\lambda_i / \mu \gamma d_i}$$

Multiplicando a equação (2) por λ_i / γ e somando-se sobre todos os is, obtém-se:

$$\frac{1}{\gamma}\sum_{i=1}^{m}\frac{\lambda_i}{\mu C_i-\lambda_i}=\frac{1}{\gamma}\sum_{i=1}^{m}\sqrt{\gamma\lambda_i d_i/\mu\beta}$$

Observando-se que o termo da esquerda é T então:

$$T=\frac{1}{\gamma}\sum_{i=1}^{m}\sqrt{\gamma\lambda_i d_i/\mu\beta}$$

Calculando-se β desta última equação, tem-se:

$$(4)\qquad \sqrt{\beta}=\frac{1}{T}\sum_{i=1}^{m}\sqrt{\lambda_i d_i/\mu\gamma}$$

Substituindo-se a equação (4) na equação (3):

$$C_i=\frac{\lambda_i}{\mu}+\frac{1}{T}\sum_{j=1}^{m}\sqrt{\lambda_j d_j/\mu\gamma}\sqrt{\lambda_i/\mu\gamma d_i}$$

ou

$$(5)\qquad C_i=\frac{\lambda_i}{\mu}\left[1+\frac{1}{\gamma T}\frac{\sum_{j=1}^{m}\sqrt{\lambda_j d_j}}{\sqrt{\lambda_i/d_i}}\right]$$

Substituindo-se a equação (5) na equação (1):

$$\text{Custo}=\sum_{i=1}^{m}\left[d_i\frac{\lambda_i}{\mu}+x_i\right]+\frac{1}{\gamma T}\left[\sum_{i=1}^{m}\sqrt{\lambda_i d_i/\mu}\right]^2$$

Considerando-se a rede do exemplo e a função custo com $d_i = 1$ e $x_i = 0$ e as capacidades dadas no exemplo, o atraso total é de 134 ms sendo que o custo, considerando $d_i = 1$ e $x_i = 0$, é de 280 unidades.

Podemos determinar as capacidades dos canais que produzam um atraso de 100 ms em lugar dos 134ms.

k	Linha	λ_i (pac/seg)	Ótima		Uniforme		Proporcional ao tráfego	
			C_i (kbps)	T_i (ms)	C_i (kbps)	T_i (ms)	C_i (kbps)	T_i (ms)
1	C1	15	24,3	65	21,7	82	25,1	61
2	C2	18	27,9	59	21,7	109	30,1	51
3	C3	15	24,3	65	21,7	82	25,1	61
4	C4	9	16,7	84	21,7	55	15,1	102

5	C5	8	15,4	89	21,7	52	13,4	114
6	C6	14	23,1	67	21,7	76	23,4	65
7	C7	12	20,6	73	21,7	66	20,1	76
8	R1	15	24,3	65	21,7	82	25,1	61
9	R2	18	27,9	59	21,7	109	30,1	51
10	R3	15	24,3	65	21,7	82	25,1	61
11	R4	9	16,7	84	21,7	55	15,1	102
12	R5	8	15,4	89	21,7	52	13,4	114
13	R6	14	23,1	67	21,7	76	23,4	65
14	R7	12	20,6	73	21,7	66	20,1	76

O custo com a designação ótima de capacidades será de 305 unidades (em lugar dos 280 originais) garantindo o atraso médio de 100 ms (em lugar de 134 ms).

A designação uniforme consiste em distribuir as capacidades uniformemente entre os canais de forma que o custo continue 305 unidades. Neste caso, o atraso médio total é 114,5ms.

A designação de capacidade proporcional ao tráfego do canal (λ_i), limitado ao custo de 304 unidades, resulta em atraso médio total de 101,7 ms.

	Custo	Atraso médio total (ms)
Ótima	305	100,0
Uniforme	305	114,5
Proporcional	305	101,7

Neste exemplo em que a rede está com pouco tráfego a vantagem da designação ótima não é tão evidente.

Quando se coloca mais tráfego na rede, a designação ótima fica muito mais vantajosa que as demais.

5 Bibliografia

[1] Jain, R., “The Art of Computer Systems Performance Analysis”, John Wiley & Sons Inc, ISBN: 0-471-50336-3, 1991, 685 p.

[2] Cassandras, C. G., ”Discrete Event Systems: Modeling and Performance Analysis”, Aksen Associates Incorporated Publishers, 1993 , ISBN: 0-256-11212-6, 790p.

[3] Menascé, D. A., Almeida, V. A. F., “Scaling E-Business: Technologies, Models, Performance and Capacity Planning”, Prentice-Hall, ISBN: 0-13-086328-9, 2000, 449p.

6 Exercícios

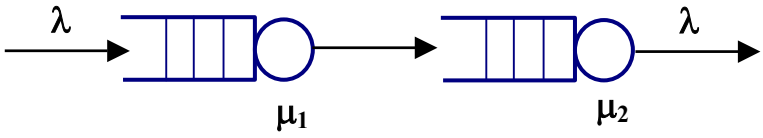
1) Durante um período de observação de 10 segundos, 400 pacotes foram atendidos por um gateway cuja UCP pode atender 200 pacotes por segundo. Qual a utilização da UCP do gateway?
Resp.: 0,20.

2) Em um sistema de time-sharing foi observada uma vazão de 5 programas por segundo em um período de observação de 10 minutos. Se o número médio de programas no sistema durante este período é 4, qual o tempo médio de resposta?
Resp.: 4/5 s.

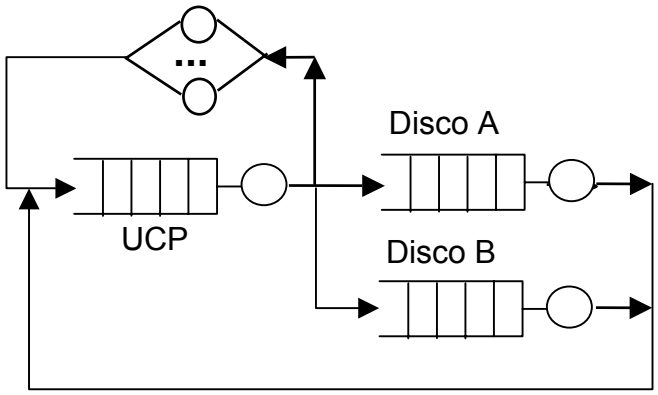
3) Durante um período de observação de 10 segundos, 40 requisições foram atendidas por um servidor de arquivos. Cada requisição requer dois acessos a disco. O tempo médio de serviço do disco é 30 ms. Qual é a utilização média do disco durante este período?
Resp.: 0,24.

4) Um sistema distribuído tem um servidor de impressão com velocidade de impressão de 60 páginas por minuto. Observou-se que o servidor imprimiu 500 páginas em um período de observação de 10 minutos. Se cada programa imprime em média 5 páginas, qual é a taxa de execução de programas neste sistema?
Resp.: 10 prog/min.

5) Um sistema consiste de duas filas encadeadas, sendo que a taxa de chegada ao sistema é λ e a taxa de serviço de cada fila é respectivamente μ_1 e μ_2. Considerando que o sistema é balanceado, construa o diagrama de estados deste sistema em que cada estado é indicado como n_1n_2 onde n_i é o número de usuários na fila i. Como você calcularia a probabilidade do sistema estar no estado n_1n_2 isto é, de existirem n_1 usuários na fila 1 e n_2 usuários na fila 2?



6) Em um sistema de timesharing com 2 discos (para usuários e sistema), após o uso da UCP, a probabilidade de um programa utilizar o disco A é de 0,80, de utilizar o disco B é de 0,16 e de utilizar os terminais é de 0,04. O tempo que o usuário fica pensando é de 5 segundos, os tempos de serviço dos discos A e B são de 30 e 25 ms respectivamente, e o tempo médio de serviço por visita à UCP é 40 mseg. Considerando que com 20 usuários a utilização do disco A é 60%, realize a análise do sistema.



a) Para cada programa, qual é a taxa de visitas à UCP, disco A e disco B?
b) Para cada dispositivo, qual é a demanda total de serviço?
c) Qual é a utilização da UCP e do disco B?
d) Qual é o tempo médio de resposta?
e) Qual dispositivo é o gargalo do sistema?
f) Qual é o tempo de resposta mínimo do sistema (independente do número de usuários)?
g) Qual é a utilização máxima do disco A (independente do número de usuários)?
h) Qual é a vazão máxima deste sistema (independente do número de usuários)?
i) Qual mudança na velocidade da UCP é recomendada para obter-se um tempo de resposta de 10 segundos com 25 usuários? Também serão necessários discos A e B mais rápidos?
j) Escreva as expressões para os limites assintóticos da vazão e do tempo de resposta e desenhe os gráficos correspondentes.

Resp.:a) 25; 20 e 4 b) 1; 0,6 e 0,1 c) 1 e 0,1 d) 15 s e) UCP f) 1,7 g) 0,60 h) 1 prog/s i) $R \geq \max\{D, N D_{max} - Z\} \Rightarrow D_{max} \leq 0.6$ e UCP 40% mais rápida j) $X \leq \min\{ N / 6.7 ; 1\}$; $R \geq \max\{1.7, N-5\}$.

7) No exercício anterior, qual dispositivo será o gargalo do sistema nas seguintes situações:
a) A UCP for substituída por outra duas vezes mais rápida?
b) O disco A for substituído por outro duas vezes mais lento?
c) O disco B for substituído por outro duas vezes mais lento?
d) O tamanho de memória for reduzido de forma que os programas façam 20 vezes mais visitas ao disco B devido ao aumento das falhas de páginas?

Resp.: a) disco A; b) disco A; c) UCP; d) disco B.

8) Três linhas de comunicação paralelas ligam os nós A e B. As três linhas transportam 5, 10 e 15 pacotes por segundo respectivamente. O tamanho médio do pacote é 1000 bits. A capacidade total de 60 Kbps está disponível para alocar às três linhas. Calcule o atraso para:
a) Atribuição de capacidade uniforme
b) Atribuição de capacidade proporcional ao tráfego
c) Atribuição ótima de capacidade.

9) Dada a rede definida pela tabela abaixo que também inclui o tráfego fim-a-fim entre pares de nós, determine o número médio de pulos por pacote, o atraso médio na rede e o fator de escala máximo que a rede suporta. O tráfego é completo, isto é, não existe tráfego reverso. Cada link é Full-duplex e tem capacidade de 50 Kbps e o tamanho médio do pacote é 1000 bits. Considere as distribuições exponenciais.

Enlace	Caminho	Tráfego em kbps	Enlace	Caminho	Tráfego em kbps
AB		10	DH	DGFH	6
AC	ABC	14	CI	CDEI	10
AD	ABCD	20	GB	GFB	8
BG	BFG	6	HD	HFGD	14
BE	BCDE	8	HE	HIE	3
BI	BFHI	6	IA	IHA	10
CE	CDE	10	IC	IEDC	15

10) Se o tamanho do pacote for aumentado para 2000 bits, como será afetada a média do número de pulos, o tempo médio de atraso e fator máximo de escala?