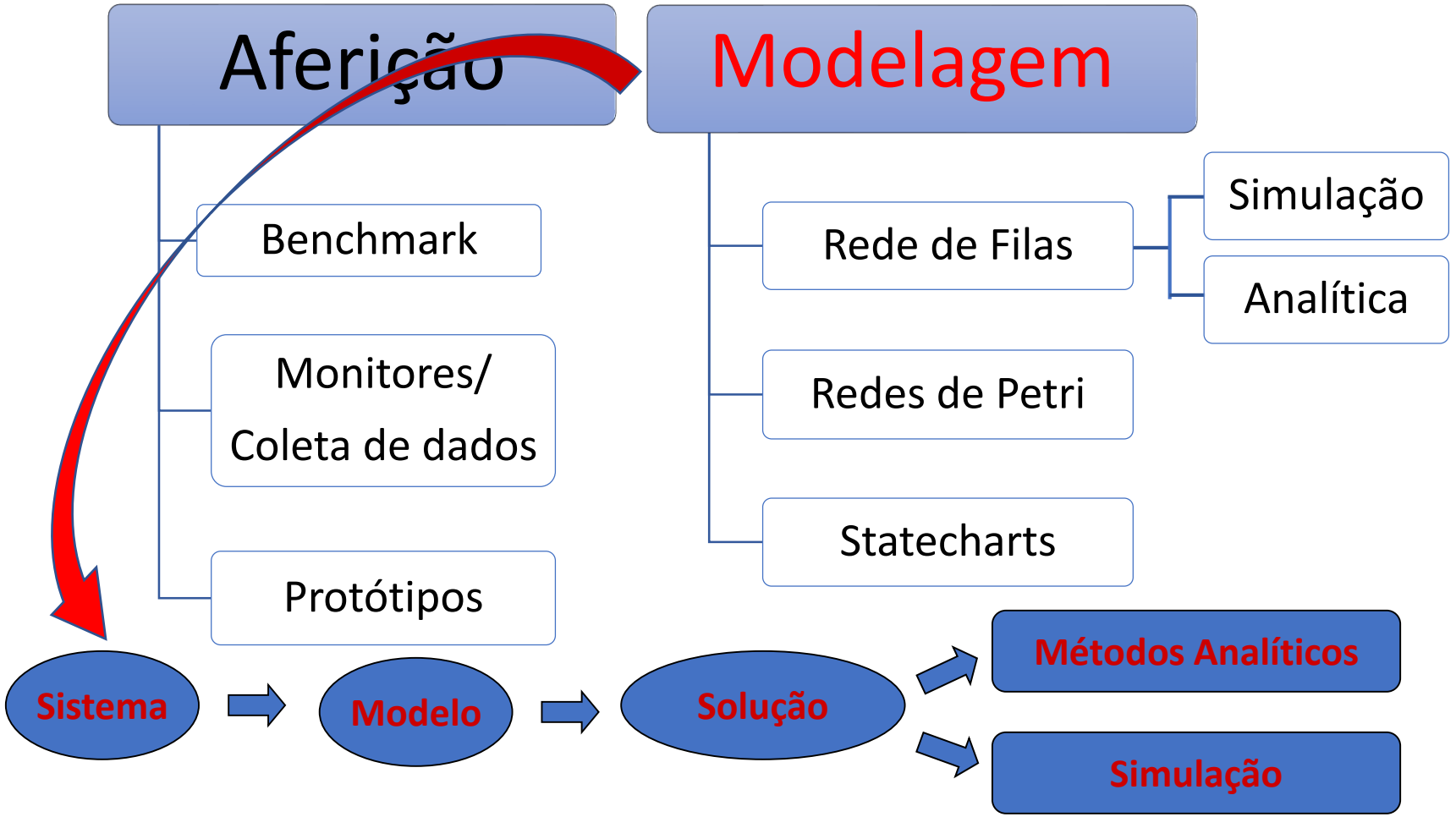
SSC0643 e SSC5872
Avaliação de Desempenho de Sistemas Computacionais

11ª Aula – Modelagem Analítica

Profa. Sarita Mazzini Bruschi

sarita@icmc.usp.br

Técnicas de Avaliação de Desempenho

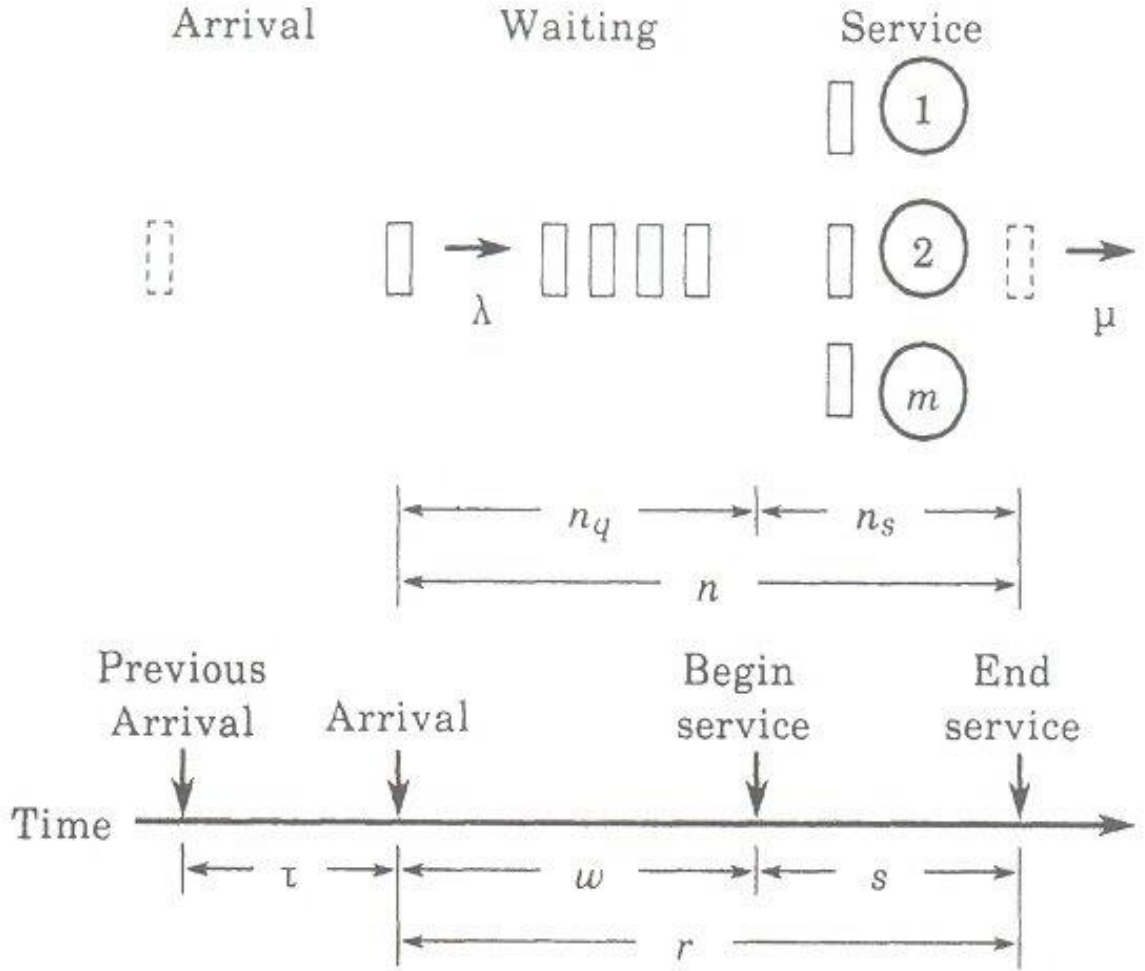


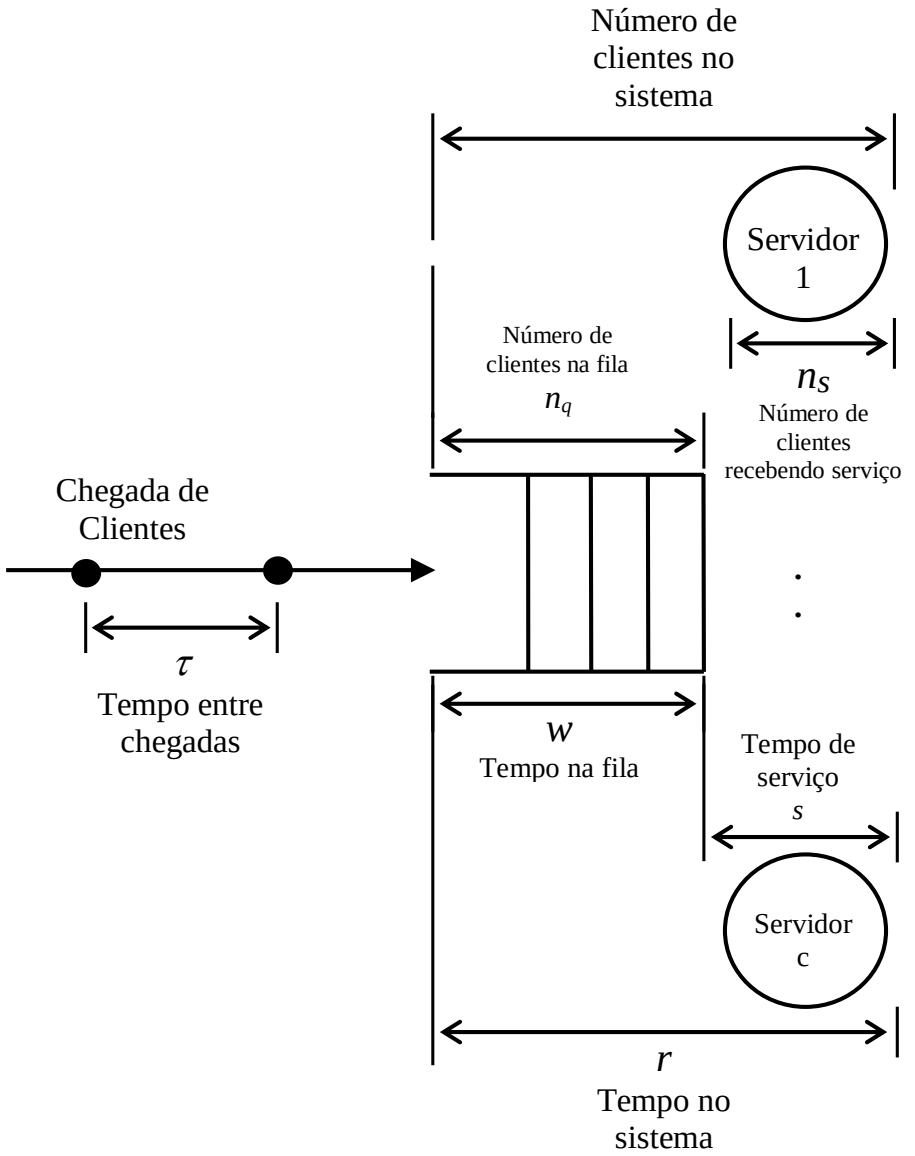
Modelagem Analítica

- Definição de modelos analíticos
  - Conjunto de fórmulas e/ou algoritmos computacionais que fornecem os valores das medidas de desempenho em função de um conjunto de parâmetros de entrada de carga de trabalho
  - De modo a serem tratáveis, normalmente são menos detalhados do que os modelos de simulação, o que implica que eles tendem a ser menos precisos mas mais eficientes em termos de execução

Modelagem Analítica

- Baseado na Teoria de Rede de Filas
- As principais características dessa modelagem já foram descritas nos slides de Simulação
- Agora estudaremos como resolver analiticamente esse modelo utilizando algumas Leis Fundamentais


Arrival
Waiting
Service
1
2
m
λ
μ
n_q
n_s
n
Previous Arrival
Arrival
Begin service
End service
Time
τ
w
s
r


Número de clientes no sistema
Servidor 1
Número de clientes na fila
n_q
n_s
Número de clientes recebendo serviço
Chegada de Clientes
τ
Tempo entre chegadas
w
Tempo na fila
Tempo de serviço
s
Servidor c
r
Tempo no sistema

Algumas definições básicas

- Período de observação: T
- Número de chegadas de clientes observadas em T : A_i
- Número de saídas/términos de clientes observadas em T : C_i
- Tempo que o sistema ficou ocupado processando: B_i

Valores de entrada dos modelos

- Taxa de chegada: $\lambda_i \equiv \frac{número\ de\ chegadas}{tempo} = \frac{A_i}{T}$
- Throughput: $X_i \equiv \frac{número\ de\ términos}{tempo} = \frac{C_i}{T}$
- Utilização: $U_i \equiv \frac{tempo\ ocupado}{tempo} = \frac{B_i}{T}$
- Tempo médio de serviço: $S_i \equiv \frac{tempo\ ocupado}{número\ de\ términos} = \frac{B_i}{C_i}$

- Em um sistema em equilíbrio, Taxa de chegada = *Throughput*
- Não confundir o valor λ_i com o valor definido como média na distribuição de probabilidade da taxa de chegada. Em uma distribuição exponencial, a média (valor esperado) é $1/\lambda$, ou seja, $\lambda = 1/média$

Lei de Little

- Relaciona o número de clientes no sistema com o tempo médio dispendido no sistema
- Provada pela primeira vez por LITTLE (1961), aplica-se a qualquer sistema ou parte do mesmo no qual o número de usuários que entram no sistema é igual ao número de usuários que terminam o atendimento.
- Número médio de clientes no sistema = taxa de chegada x tempo médio de resposta

$$Q_i = \lambda_i \times R_i$$

- $R_i = S_i + W_i$ (tempo médio de resposta = tempo médio de serviço + tempo médio de espera)
- Se o sistema está em equilíbrio, $\lambda_i = X_i$, portanto $Q_i = X_i \times R_i$
- Pode-se aplicar a lei de Little a qualquer sistema ou subsistema (caixa preta)

Lei de Little

- Exemplo: Um servidor de arquivos NFS foi monitorado durante 30 minutos e o número observado de operações de I/O foi 10.800. Apurou-se que o número médio de pedidos ativos no NFS era três. Qual o tempo médio de resposta por pedido no servidor?
- Solução:
  - $X = \frac{10800}{1800} = 6$ operações I/O por segundo
  - $Q = 3$
  - $R = \frac{Q}{X} = \frac{3}{6} = 0{,}5 \text{ segundo}$

Lei de Little

- A Lei de Little pode ser aplicada em vários níveis do sistema

- Exemplo: aplicação da Lei de Little em 3 níveis de um sistema que possui 1 CPU, 3 discos e carga balanceada (probabilidade de requisição ser enviada para cada disco é 1/3)

Lei de Little

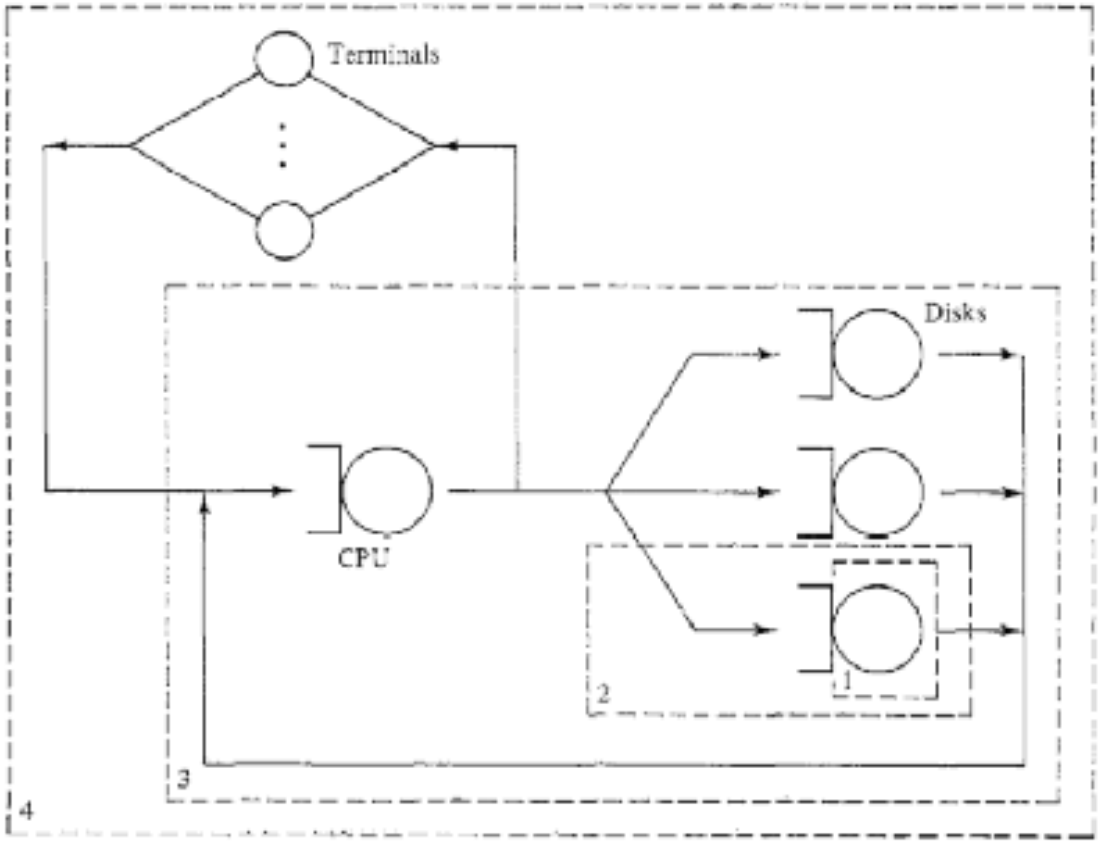


Figure 3.3 – Little's Law Applied at Four Levels

Lei de Little
Um recurso sem a fila

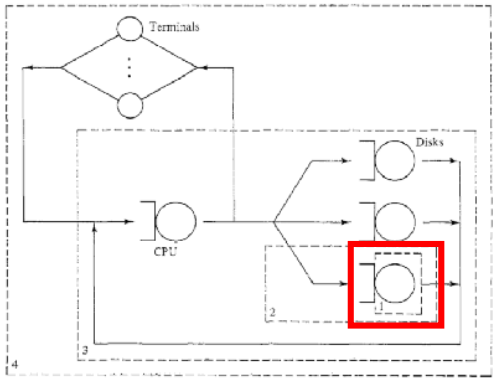


Figure 3.3 – Little's Law Applied at Four Levels

- Suponha que um disco sirva em média 40 requisições/seg e que uma requisição típica demande 0.0225 segundos para ser servida pelo disco. Qual a utilização do disco?
  - População de clientes N = U - Utilização do recurso (0 - 1)
  - Tempo de residência R = S - requisito (tempo) de serviço médio por cliente (não inclui atraso na fila)
  - $R = S = 0{,}0225$
  - $X = 40$
  - $N = U = XS = 40 \times 0{,}0225 = 0{,}9$
  - $U = 90\%$

Lei de Little
Um recurso sem a fila

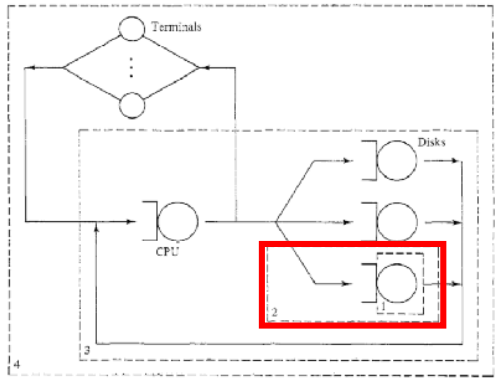


Figure 3.3 – Little's Law Applied at Four Levels

- Exercício: Suponha que para o mesmo disco do exemplo anterior, foi verificado que existem, em média 4 requisições para leitura pendentes. Qual o tempo que uma requisição permanece na fila de espera do disco? Qual o tamanho médio da fila de espera?
  - $N = 4$
  - $X = 40$
  - $N = XR \rightarrow R = \frac{N}{X} = \frac{4}{40} = 0{,}1\ seg$
  - Tempo na fila: $R\ - 0{,}0225 = 0{,}1\ - 0{,}0225 = 0{,}0775\ seg$
  - Tamanho da fila: $N - U = 4 - 0{,}9 = 3{,}1$

Lei de Little
Um recurso sem a fila

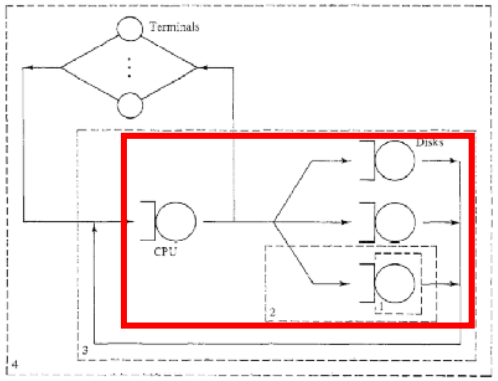


Figure 3.3 – Little's Law Applied at Four Levels

- Exercício: Suponha que o servidor de arquivos consiga processar em média 1 requisição a cada 2 segundos e que haja em média 7.5 usuários submetendo requisições simultaneamente. Qual o tempo de resposta médio observado por estes usuários?
  - N = interações a nível de cliente (realmente clientes)
  - X = taxa de interações entre terminais e sistemas
  - R = tempo de resposta
  - $N = 7{,}5 \qquad X = 0{,}5$
  - $N = RX \qquad R = \frac{N}{X} = 15\ seg$

Lei da Utilização

$$U_i = \frac{B_i}{T} = \frac{C_i}{T} \times \frac{B_i}{C_i}$$

$$U_i = X_i \times S_i$$

- Exemplo: Considere um roteador em que os pacotes chegam a uma taxa de 125 pps e o roteador leva em média 2 ms para encaminhá-los. Qual a utilização do sistema?
- $X_i = taxa\ de\ saída(throughput) = \ taxa\ de\ chegada = 125pps$
- $S_i = 2ms = \ 0{,}002\ segundos$

$$U_i = X_i \times S_i = 125 \times 0{,}002 = 0{,}25 = 25\%$$

Lei do Fluxo Forçado

- Relaciona o *throughput* do sistema com o *throughput* dos dispositivos individuais
- Se o período de observação *T* for tal que o número de chegadas em cada dispositivo é igual ao número de saídas, i.e., $A_i = C_i$, diz-se que o dispositivo satisfaz a *Hipótese de Balanço* (*job flow balance*)
- Seja V_i o número médio de visitas ao recurso i por uma tarefa
- Cada pedido que termina precisa passar, em média, V_i vezes pelo recurso i. Assim, se X pedidos foram concluídos por unidade de tempo (*throughput* do sistema), temos que $V_i X$ pedidos terão passados pelo recurso i:

$$X_i = V_i X$$

Lei do Fluxo Forçado

- Exemplo: Suponha que cada transação finalizada em um sistema de servidor de banco de dados precisa, na média, de duas operações de E/S no disco. Isso significa que cada requisição ao servidor de banco de dados finalizada visita o disco duas vezes. Qual o *throughput* do disco em termos de operações de E/S dado que o *throughput* do sistema é 3,8 transações completadas por segundo?
- $X_i = V_i X = 2 \times 3{,}8 = 7{,}6\ operações\ de\ E/S$

Lei da Demanda de Serviço

- A lei da demanda de serviço é associada tanto com o recurso como com um conjunto de requisições que usam o recurso
- A demanda de serviço (D_i) é definida como a soma total de tempo gasto por uma requisição de um determinado tipo obtendo o serviço de um recurso i
- Durante sua passagem pelo sistema, uma requisição pode visitar diversos dispositivos, possivelmente muitas vezes
- Para qualquer requisição, sua demanda por serviço é a soma de todos os tempos de serviço durante todas as visitas a um determinado recurso

Lei da Demanda de Serviço

- Combina a Lei da Utilização e a do Fluxo Forçado
- $U_i = X_i S_i = X V_i S_i$ ou $U_i = X D_i$

- Onde $D_i = V_i S_i$ é a demanda total de serviço do i-ésimo dispositivo

- O dispositivo com maior demanda de serviço tem a maior utilização e pode tornar-se o *gargalo do sistema*

Lei da Demanda de Serviço

- Exemplo: As transações de um banco de dados realizam uma média de 4,5 operações de I/O no servidor de BD. O servidor foi monitorado durante uma hora e, durante esse período, 7.200 transações foram concluídas.
  a) Qual a taxa média de processamento no disco?

$$X = \frac{7200}{3600} = 2 \text{ transações / seg}$$

$$X_{disco} = X \times V_{disco} = 4{,}5 \times 2 = 9 \text{ operações de I/O / seg}$$

b) Se cada I/O de disco leva 20 ms em média, qual a utilização do disco?

$$U_{disco} = X_{disco} \times S_{disco} = 9 \times 0{,}02 = 0{,}18 = 18\%$$

c) Qual a demanda de serviço do disco?

$$D_{disco} = V_{disco} \times S_{disco} = 4{,}5 \times 0{,}02 = 0{,}09\ seg\ (90\ ms)$$

Lei Geral do Tempo de Resposta

- Sistemas de tempo compartilhado podem ser divididos em dois subsistemas: o subsistema de terminais e o subsistema central de processamento
- Dados os comprimentos individuais $Q_i$das filas de cada dispositivo, podemos calcular Q:

$$Q = Q_1 + Q_2 \not\geq \cdots + Q_M$$

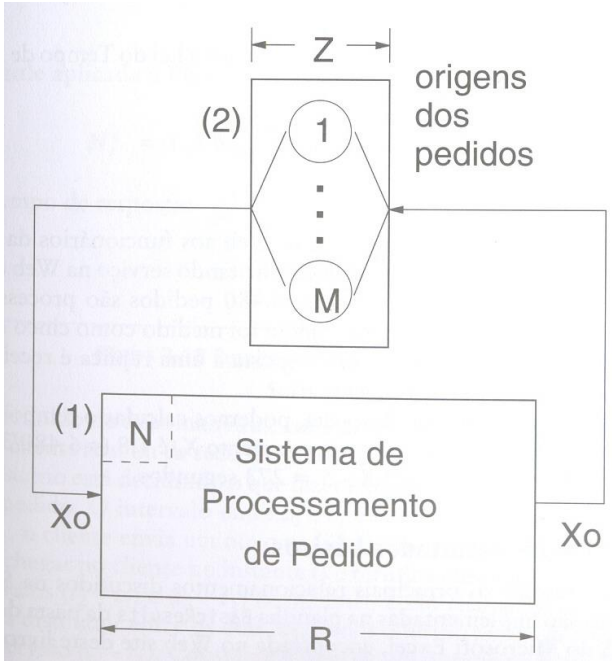
$$XR = X_1R_1 + X_2R_2 \not\geq \cdots + X_MR_M$$

- Dividindo ambos os lados por X e aplicando a lei de Little, temos

$$R = V_1R_1 + V_2R_2 \not\geq \cdots + V_MR_M \text{ ou } R = \sum_{i=1}^{M} R_iV_i$$

Lei Geral do Tempo de Resposta

- Num sistema interativo, os usuários geram pedidos que são processados pelo subsistema central e os resultados voltam ao terminal. Após um tempo ocioso Z, o usuário submete o próximo pedido.
- Seja $\overline{M}$ e $\overline{N}$ o número de clientes que estão "pensando" (em suas próprias estações de trabalho) e esperando uma resposta do servidor



Lei Geral do Tempo de Resposta

- $M = \bar{M} + \bar{N}$
- Aplicando-se a lei de Little ao subsistema central, temos:

$$\bar{N} = XR$$

- Agora, aplicando-se a lei de Little nos M terminais

$$\bar{M} = XZ$$

- Considerando que um cliente ou está sendo atendido ou está ocioso:

$$M = \bar{N} + \bar{M} = XR + XZ = X(R + Z)$$

$$R = \frac{M}{X} - Z$$

Lei Geral do Tempo de Resposta

- Exemplo: Suponha que 10 usuários utilizem o sistema. Estes usuários fazem processamento local por 5 segundos, em média, antes de submeterem requisições ao servidor central. O tempo de resposta médio observado por eles é de 15 segundos. Qual o *throughput* do sistema?
  - M = número total de usuários
  - X = taxa de interações entre terminais e subsistema
  - Tempo de residência = tempo de resposta (R) + *think time* (Z)

  - M = 10 R = 15 Z = 5
  - $M = X(R + Z)$ $X = \frac{M}{R+Z} = \frac{10}{20} = 0{,}5\ interações/seg$

Exercícios

- Determine o tempo médio de resposta de um sistema interativo com as seguintes características conhecidas:
  - 25 terminais
  - *Think time* médio de 18 segundos
  - Cada interação faz 20 acessos ao disco, em média
  - Disco está ocupado em média 30% do tempo, durante medição
  - Tempo de serviço médio por acesso ao disco igual a 25mseg

Exercícios

- Suponha que um servidor com três discos atendendo a uma comunidade fechada tenha memória limitada: pode ocorrer swapping e, portanto, antes de competir pelos recursos do sistema central, uma interação deve competir por uma partição da memória. O sistema foi observado e medido:
  - número médio de usuários : 23
  - tempo de resposta médio percebido por um usuário: 30 s
  - *throughput* do servidor : 0.45 interações / s
  - número médio de requisições ocupando memória: 1.9
  - demanda média por CPU para cada interação: 0.63 s

Exercícios

- Responda:
  - Qual o *think time* médio de um usuário?
  - Em média, quantos usuários estão tentando obter serviço (não estão em think time)?
  - Em média, quantos usuários estão esperando na fila da memória?
  - Em média, quanto tempo se passa desde a aquisição da memória até o término de uma interação?
  - Qual o tempo médio gasto na fila da memória?
  - Qual a utilização de CPU pela carga de timesharing?
  - Qual o gargalo do sistema?