

Modelagem Analítica

Profa. Jussara M. Almeida
1º Semestre de 2014

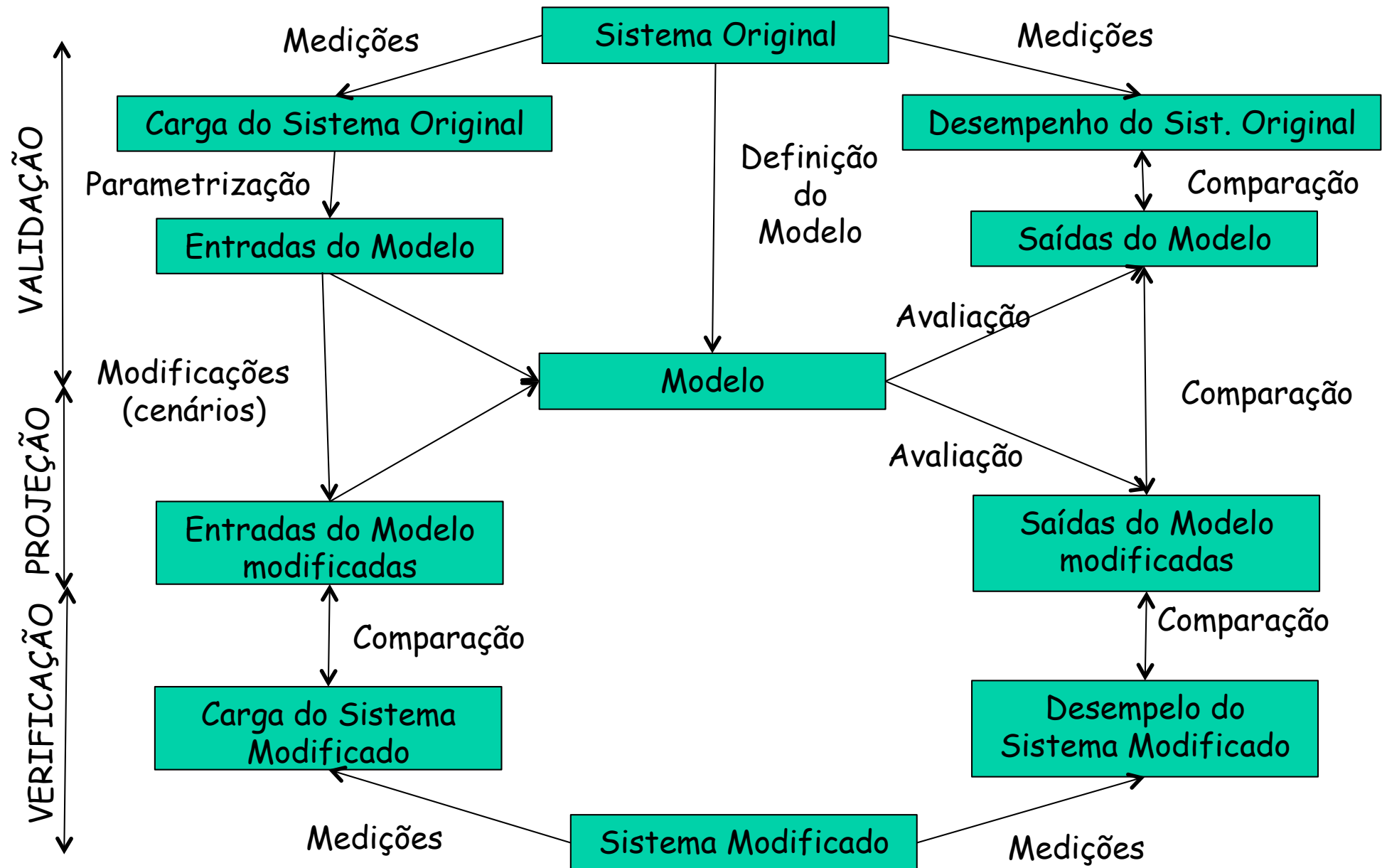
Modelagem Analítica

Um modelo é uma abstração de um sistema que captura, dentre os inúmeros detalhes do sistema, aqueles que são essenciais para o seu comportamento

Conduzindo um Estudo de Modelagem

1. Desenvolvimento do Modelo
2. Validação
3. Projeção
4. Verificação

Conduzindo um Estudo de Modelagem



Exemplo de Estudo de Modelagem

Cinco sistemas fracamente acoplados.

Sistema	Tipo de CPU	Carga
1	A	batch
2	B	interativo, batch
3	B	batch
4	A	interativo, batch
5	A	batch

Pode a carga do sistema 5 ser distribuído entre os sistemas 2, 3 e 4 sem degradar o desempenho significativamente?

Exemplo de Estudo de Modelagem

Validação: sistemas 2-5 foram medidos e modelados

Projeção: nível de multiprogramação da carga batch nos sistemas 2, 3 e 4 (# jobs) foi acrescido de 27% da carga do sistema 5 (19% p/ sistema 1)

Verificação: carga do sistema 5 foi migrada para sistemas 2, 3 e 4, sistemas foram medidos, valores comparados com saídas do modelo

Exemplo de Estudo de Modelagem

perf. measure	workload component	original system	original model	modified model	modified system
CPU util.	Interativo	76%	74%	74%	72%
	batch	11%	10%	13%	13%
	total	87%	84%	87%	85%
resp. time	Interativo	5.2 secs.	4.8 secs.	5.0 secs.	5.6 secs.
t'put.	batch	28/hr.	27/hr.	35/hr.	30/hr.

Table 2.2 – The Modelling Cycle: Case Study 2, System 2

Erro: Projeção da degradação no tempo de resposta

$$\text{Projeção} = (5.0 - 4.8) / 4.8 = 4\%$$

$$\text{Degradação real} = (5.6 - 5.2) / 5.2 = 7.5\%$$

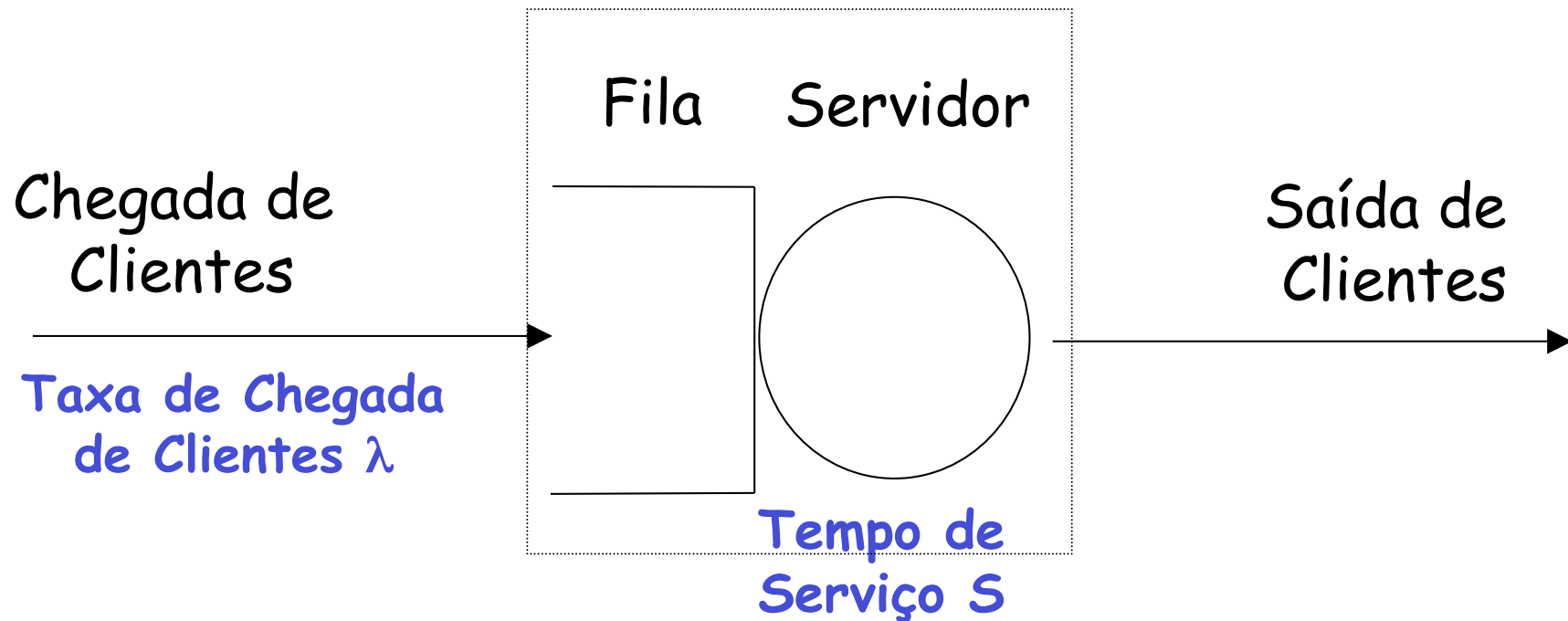
Precisão dos Resultados Analíticos

- Pergunta que não quer calar:
Meu modelo é preciso? ou
Os resultados foram validados?
- Validação de modelos de filas (típico)
Throughput e utilização: erros $\leq 5-10\%$
Tempo de resposta: erros $\leq 10-30\%$

Modelos de Redes de Filas

- Sistema é representado por uma rede de filas, que é avaliada analiticamente
- Rede de Filas é uma coleção de
 - Centros de serviços: recursos do sistema, e
 - Clientes: usuários ou transações (carga de trabalho)
- Avaliação analítica envolve a solução (possivelmente com algum software) de um conjunto de equações induzido pela rede de filas e seus parâmetros

Centro de Serviço

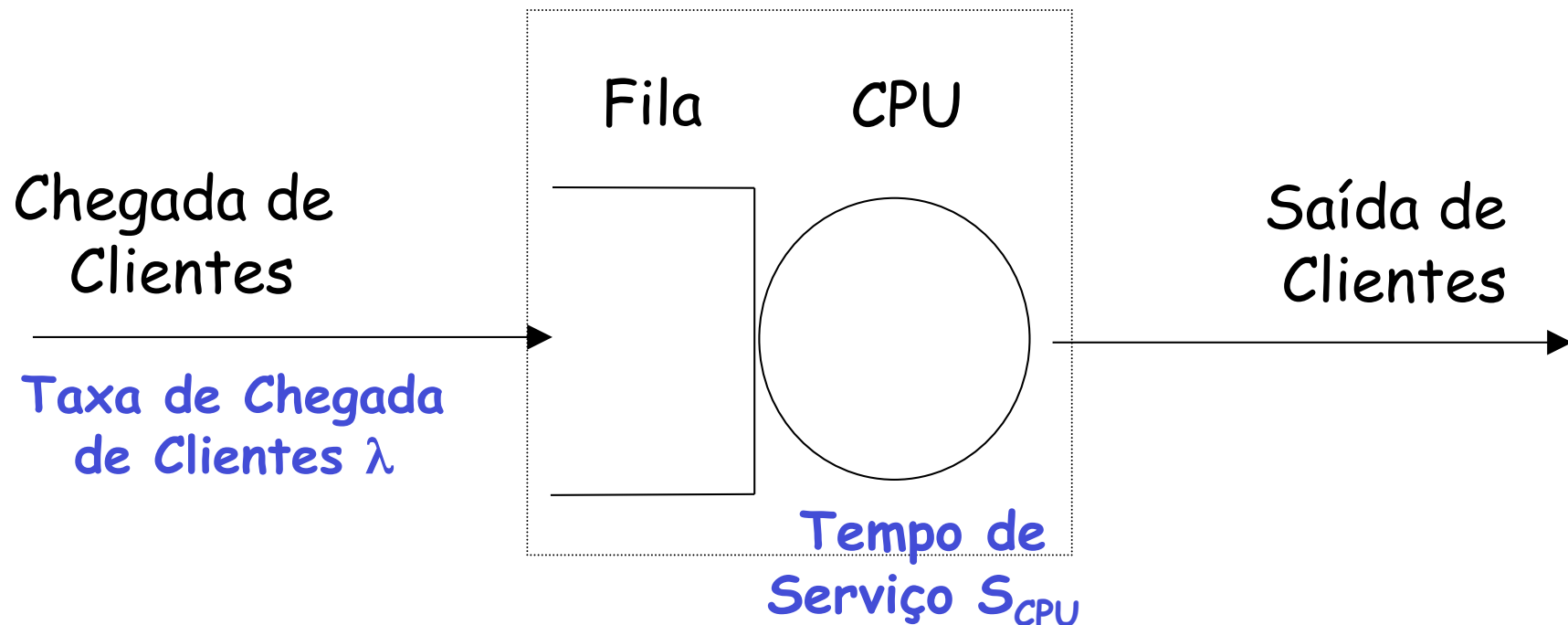


Qual o tempo de resposta (enfileiramento + serviço) médio?

Qual o número médio de clientes na fila?

Qual o tempo médio de espera por atendimento?
(tempo na fila)

Modelando CPU

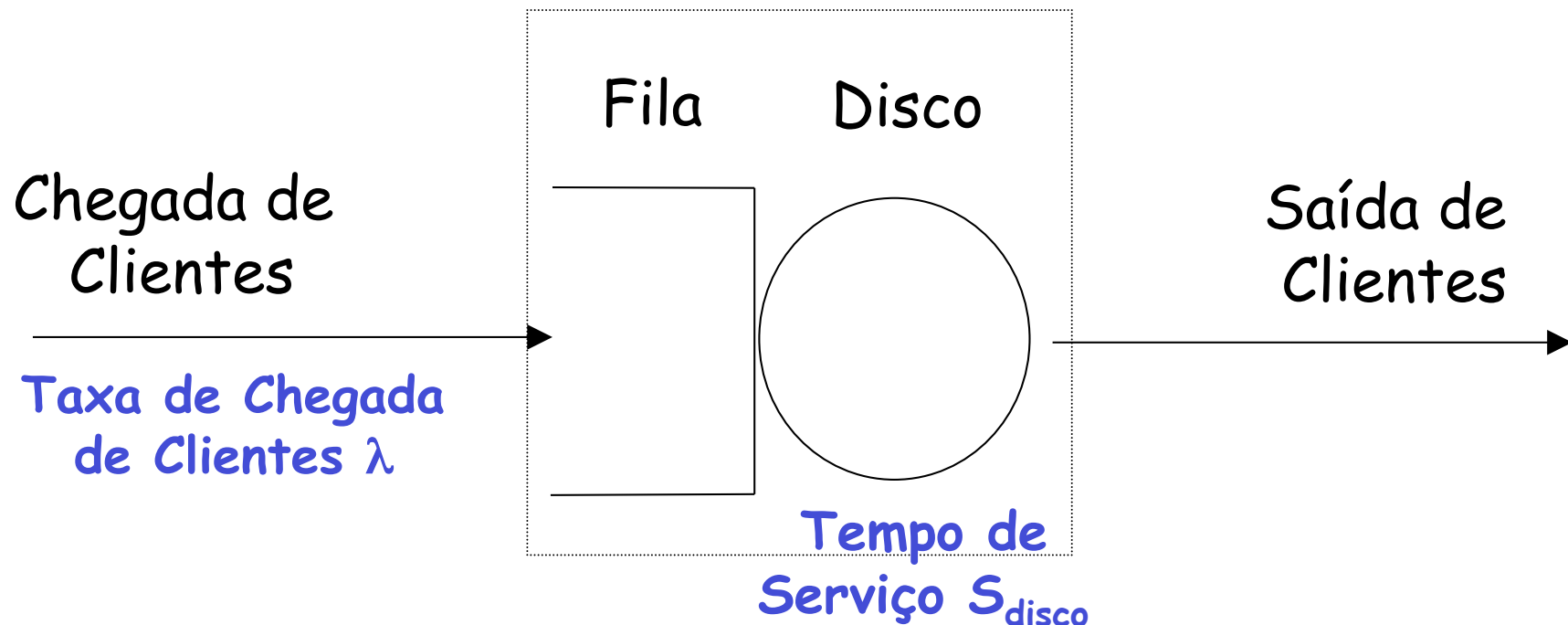


Qual a utilização da CPU (*busy time*) ?

Qual o número médio de clientes na fila de CPU?

Qual o tempo médio de espera por atendimento da CPU?
(tempo na fila)

Modelando Disco

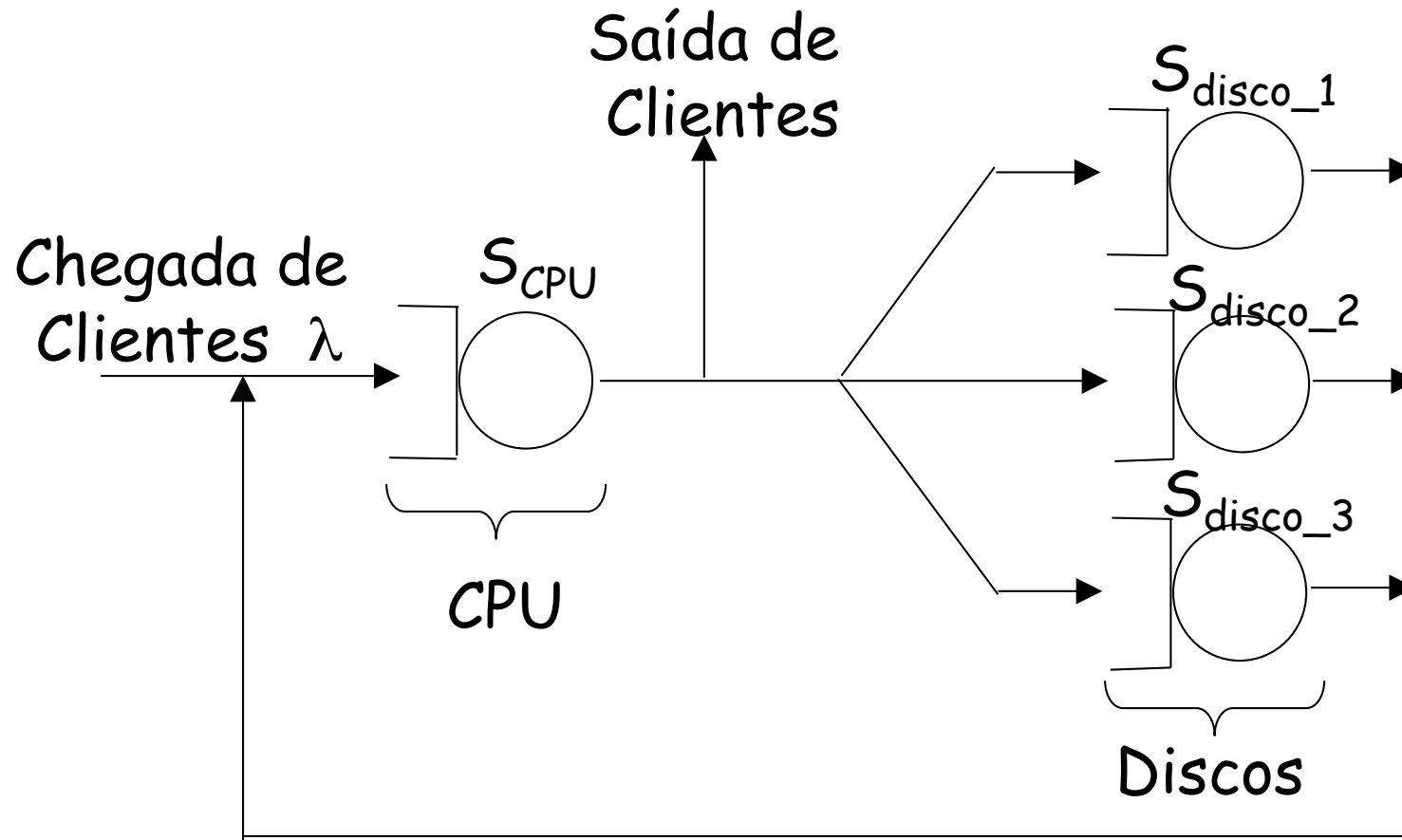


Qual o throughput do disco?

Qual o número médio de clientes na fila do disco?

Qual o tempo médio de espera por atendimento do disco?
(tempo na fila)

Múltiplos Centros de Serviço

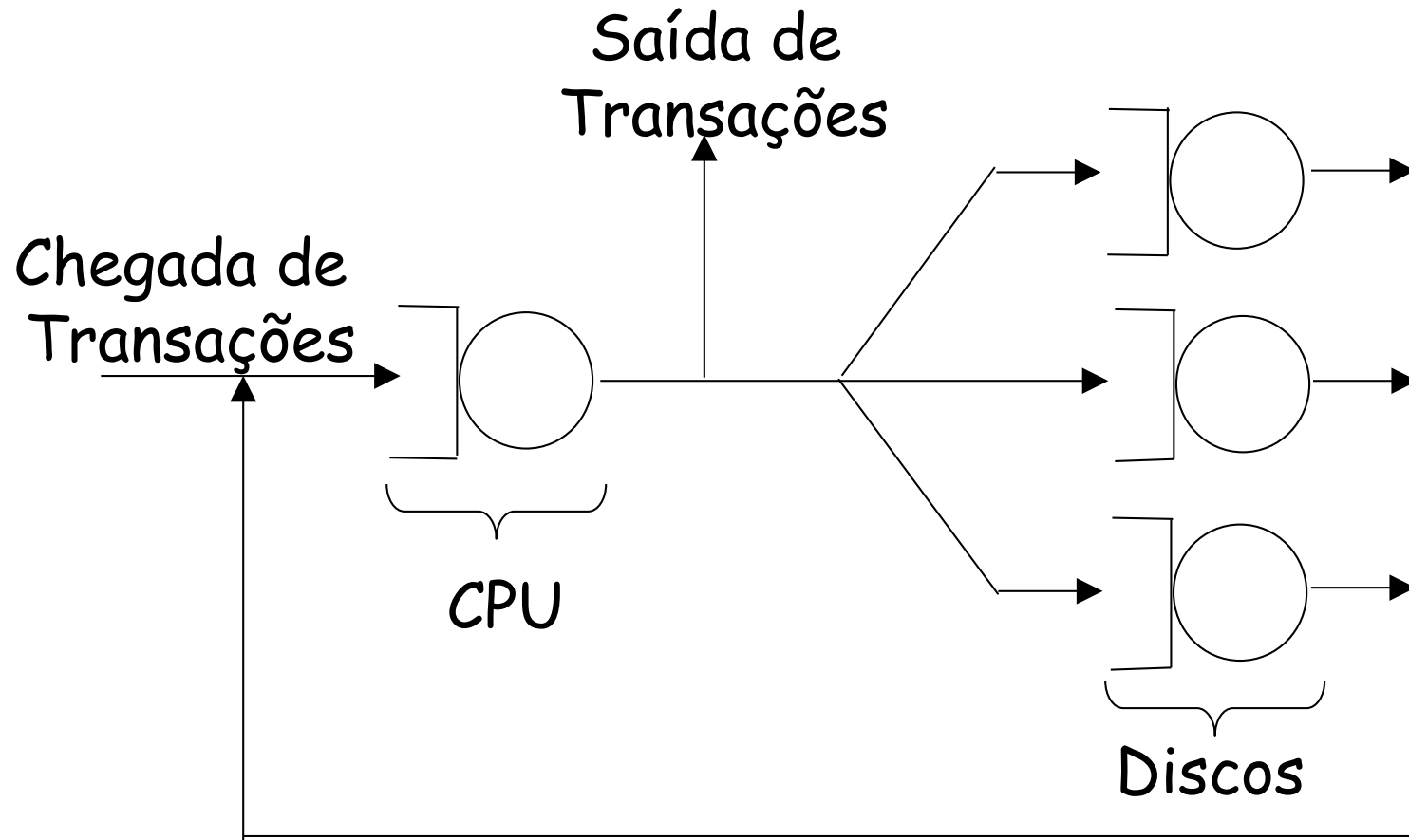


Qual o dispositivo que é o gargalo (*bottleneck*)?

Qual o tempo médio de resposta de cada requisição?

Qual o impacto no tempo de resposta se a CPU é substituída por uma 2 vezes mais rápida?

Exemplo 1: Banco de Dados



Qual o dispositivo que é o gargalo (*bottleneck*)?

Qual o tempo médio de processamento de cada transação?

Quanto tempo seria reduzido se o banco fosse migrado para uma plataforma com 4 discos do mesmo modelo do disco 3?

Exemplo 1: Banco de Dados

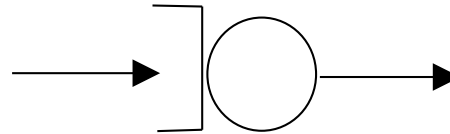
- Transações com diferentes características

Classe	% Total	Tempo médio de CPU (seg)	# médio de I/Os
Leve	45%	0.04	5.5
Médio	25%	0.18	28.9
Pesado	30%	1.20	85.0

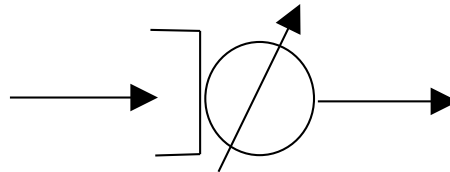
- É preciso determinar se o modelo deve ter múltiplas classes de cargas
 - Ex: demanda por serviço heterogênea, cargas de trabalho diferentes (online x batch), SLAs diferentes
- Em caso afirmativo, é preciso caracterizar a carga e parametrizar o modelo separadamente para cada classe.

Tipos de Recursos

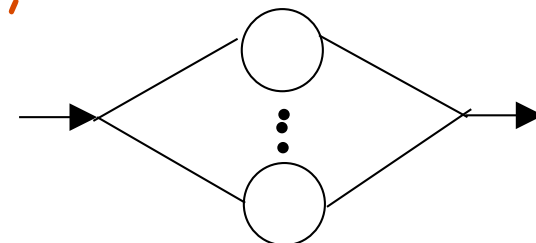
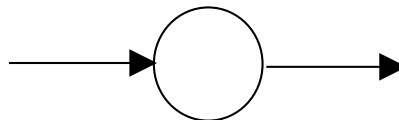
- *Load Independent*: centros de serviços com taxa de serviço constante, que não depende da carga
 - Ex: CPU, disco



- *Load Dependent*: centros de serviços cuja taxa de serviço depende do número de clientes na fila
 - Ex: fila com m servidores, LAN (colisões de pacotes), controle de congestionamento para tráfego CBR, disco (seek time)



- *Centros de atrasos*: não tem fila
 - Ex: recursos dedicados, *think time*,
recursos >> # clientes



Tipos de Classes de Cargas

- **Classes Abertas:**

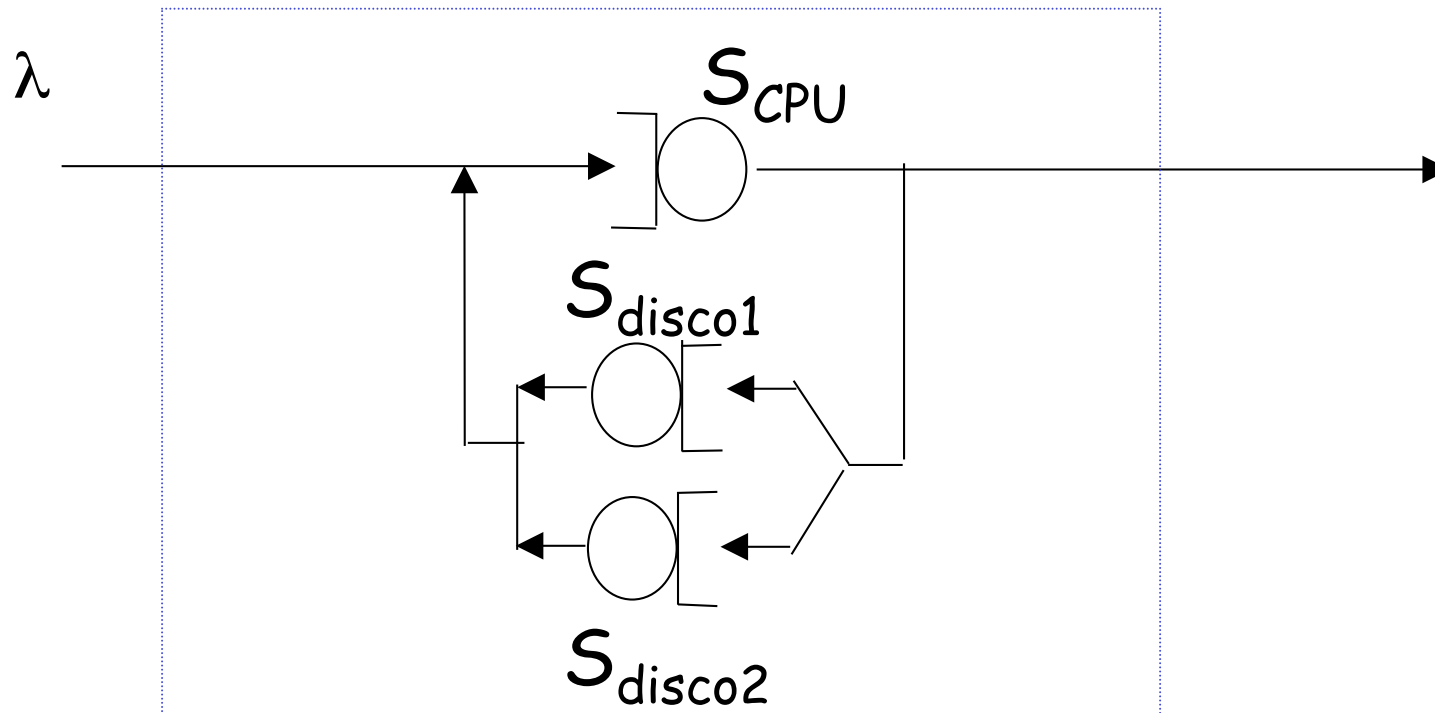
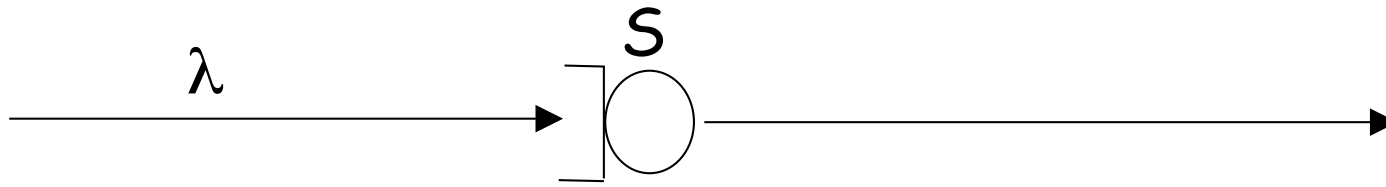
- Número ilimitado de clientes no sistema
- Ex: cargas de transações

- **Cargas Fechadas:**

- Número de clientes no sistema limitado e conhecido
- Cargas interativas e batch

Classes Abertas

- Cargas de transações

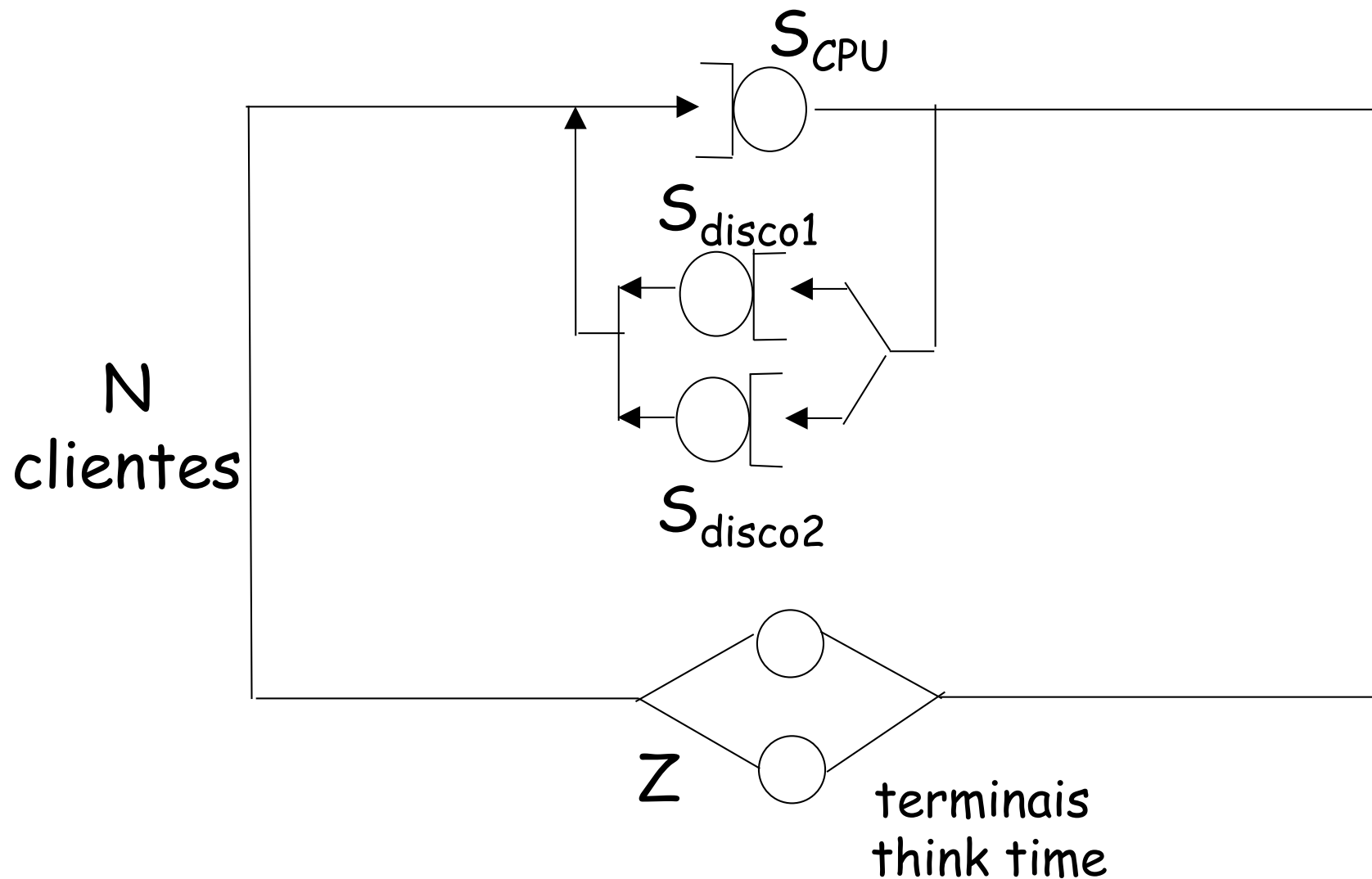


Características de Classes Abertas

- Intensidade da carga dada pela taxa de chegada
 - λ = número médio de transações que chegam por unidade de tempo,
 - independente do número de transações no sistema
- Número de clientes no sistema ilimitado
- Throughput X é parâmetro de entrada
 - $X = \lambda$ (sistema em equilíbrio)
- Classe aberta definida por
 - Taxa de chegada λ
 - Tempo de serviço em cada centro k , S_k

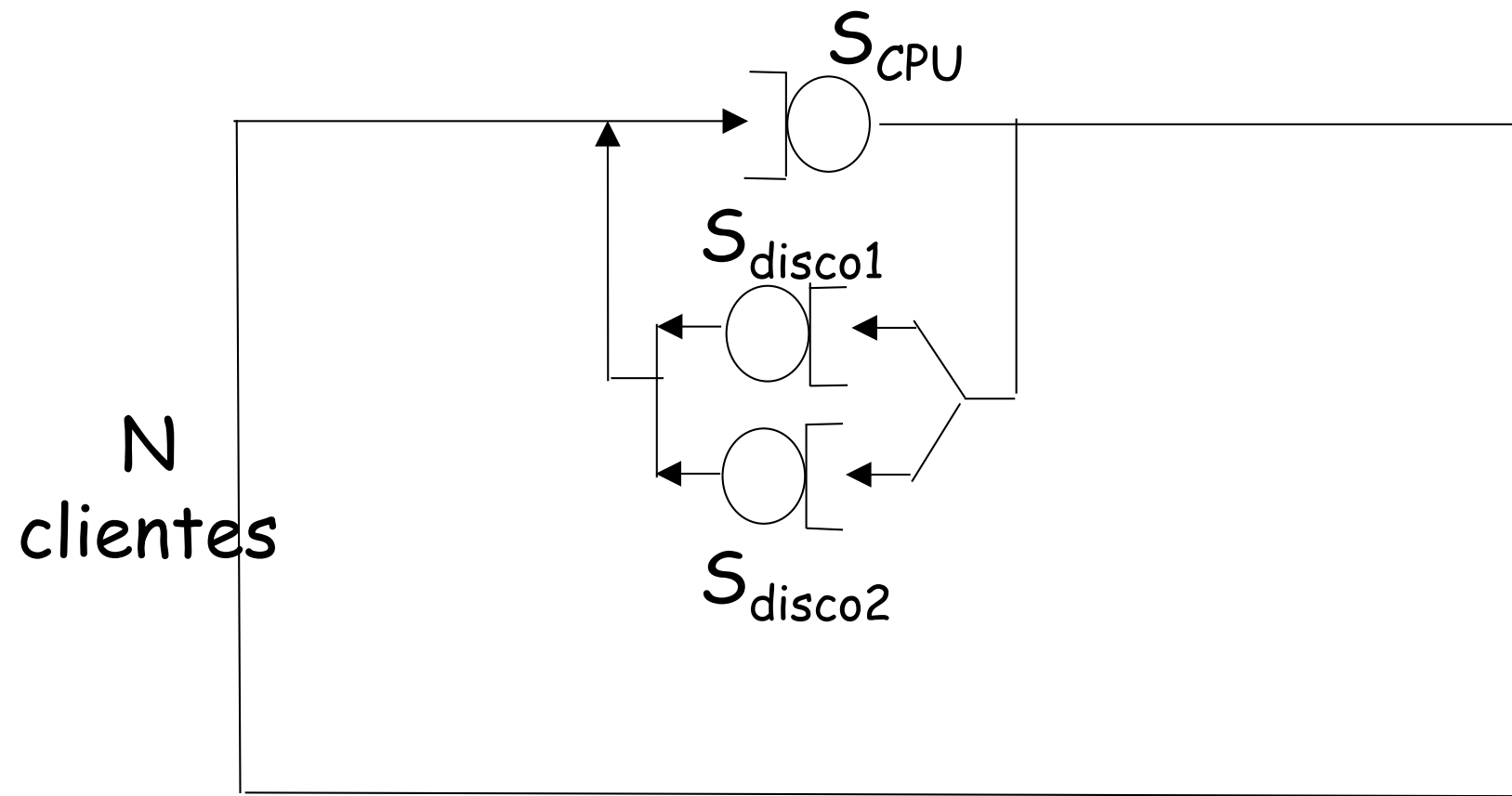
Classes Fechadas

- Cargas Interativas



Classes Fechadas

- Cargas Batch



Características de Classes Fechadas

- Intensidade da carga dada pela população de clientes
 - N = número de requisições concorrentes em execução (número de clientes)
 - Número de clientes no sistema é limitado e conhecido
- Throughput X é parâmetro de saída
- Z = think time / delay center (carga interativa)
- Classe fechada definida por
 - Taxa de chegada N
 - Tempo de serviço em cada centro k , S_k
 - Think time Z (carga interativa)

Classe Aberta ou Classe Fechada?

- Normalmente, a decisão correta é clara:
 - O número de clientes é limitado ?
- A decisão é também essencial:
 - Modelos de carga incorretos levam a resultados incorretos

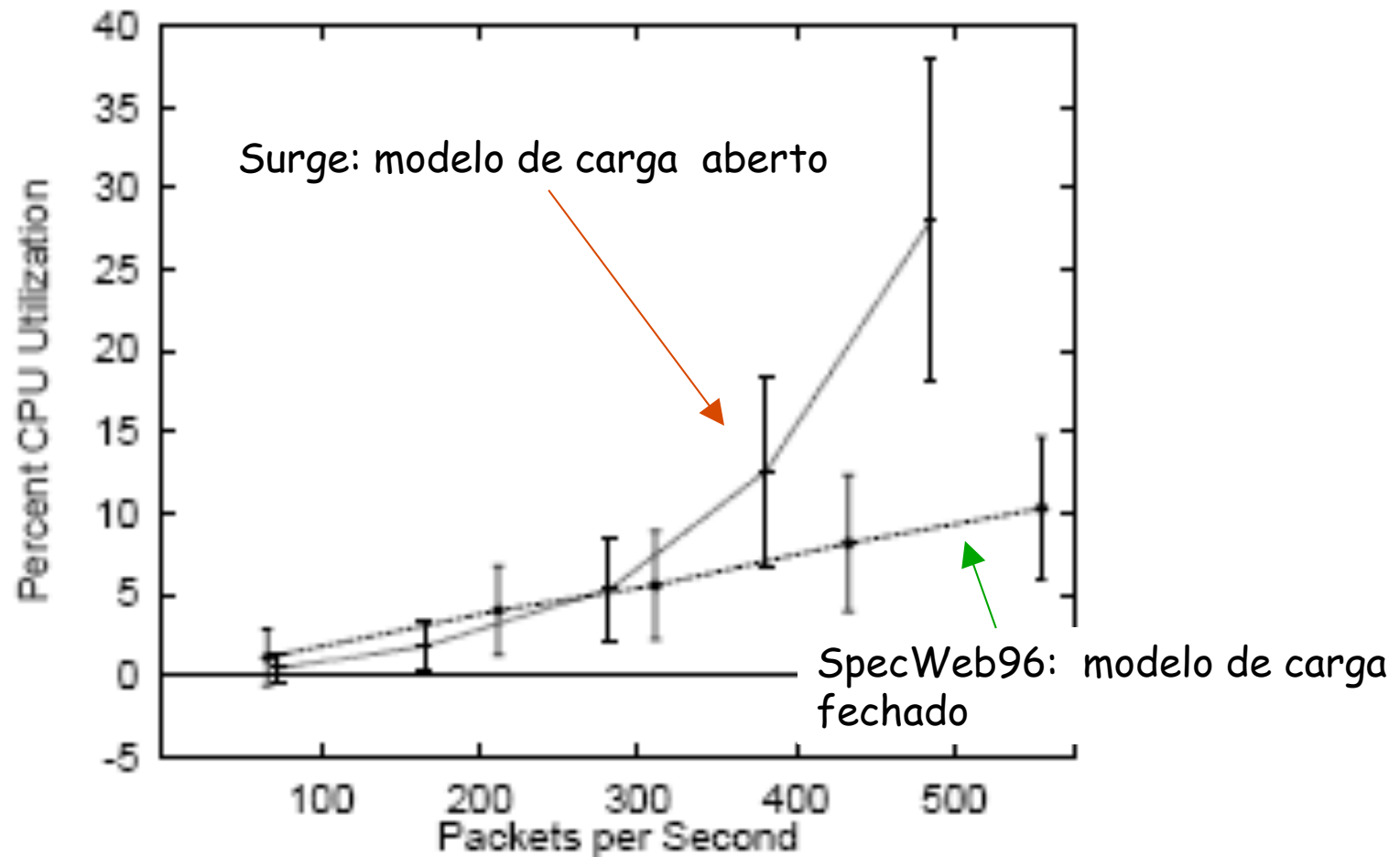
Classe Aberta ou Classe Fechada?

- Semáforo de um sistema:
- Servidor Web:
- Servidor de arquivos do DCC/UFMG:
- Memória compartilhada:

Classe Aberta ou Classe Fechada?

- Semáforo de um sistema: Fechado (N fixo)
- Servidor Web: Aberto (X fixo)
- Servidor de arquivos do DCC/UFGM: Fechado
- Memória compartilhada: Fechado

Desempenho de um Servidor Web

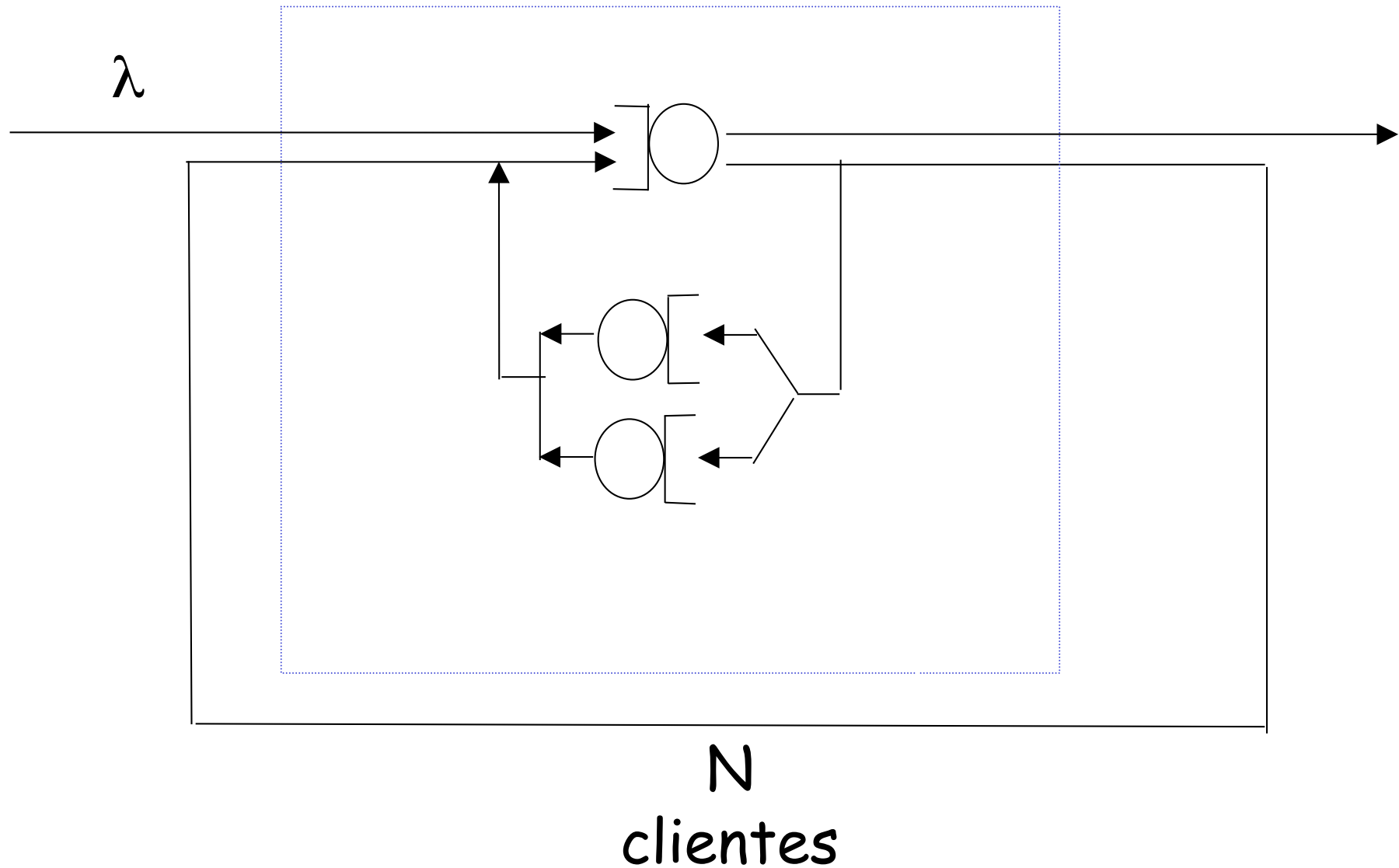


Fonte: *Generating Representative Web Workloads for Network and Server Performance Evaluation*, Paul Barford and Mark Crovella, Proc. Sigmetrics 1998

Tipos de Modelos de Filas

- **Modelos abertos:** somente classes abertas
- **Modelos fechados:** somente classes fechadas
- **Modelos mistos:** classes abertas e fechadas
 - Ex: banco de dados com processamento simultâneo de diferentes tipos de transações online e geração de relatórios de gerenciamento (batch)
 - SLAs:
 - Batch: throughput mínimo = 20 jobs/hora
 - Transações : max tempo de resposta médio
 - Transações Leves ≤ 1.2 seg
 - Transações Médias ≤ 2.5 seg
 - Transações Pesadas ≤ 8.0 seg

Exemplo de Modelo Misto



Formalizando

Notação

Modelos com Uma Classe: Entradas

- **Descrição da Carga**
 - Transação: taxa de chegada λ
 - Batch: número de clientes/jobs N
 - Interativo: número de clientes N e think time Z
 - Não é preciso definir política de escalonamento (tipicamente)
- **Descrição dos Centros de Serviço**
 - Número de centros de serviço K
 - Para cada centro k : o tipo (LI, LD, delay center)
- **Descrição das Demandas (Carga)**
 - Para cada centro de serviço k
 - Tempo de serviço S_k e número de visitas V_k ou
 - Demanda por serviço $D_k = S_k \times V_k$

Tempo de Serviço vs. Demanda por Serviço

- Tempo de serviço S_k
 - Tempo médio de serviço cada vez o cliente visita o centro de serviço k
- Demanda por serviço D_k
 - Tempo total médio que um cliente requer de serviço no centro k enquanto no sistema
- V_k : número de visitas que cada cliente faz ao centro k enquanto no sistema

$$D_k = S_k \times V_k$$

Demanda por Serviço

- Demanda por serviço D_k
 - Razão do tempo de ocupação (*busy time*) do centro B_k e o número de clientes que deixaram o sistema no período de medição C

$$D_k = S_k \times V_k = B_k / C$$

- Solução depende do produto $S_k \times V_k = D_k$ e não dos valores individuais de S_k e V_k (redes separáveis)
- Demanda total de um cliente em todos centros:

$$D = \sum_{k=1}^K D_k$$

Modelos com Uma Classe: Saídas

- **Medidas do Sistema**
 - R : Tempo médio de resposta
 - X : Throughput do sistema
 - Q : Número médio de clientes no sistema
- **Medidas por Centro**
 - U_k : utilização do centro k
 - R_k : tempo médio de residência no centro k
 - X_k : throughput do centro k
 - Q_k : tamanho médio da fila no centro k
- **Outras Medidas: Probabilidade**

Utilização de um Centro: U_k

- Proporção do tempo que o dispositivo está ocupado (prestando serviço)
- Número médio de clientes em serviço
- Ex: CPU está 70% ocupada ou em 70% do tempo no intervalo de medição, algum processo estava sendo executado.

Tamanho Médio da Fila : Q_k

- Q_k : Inclui *todos* clientes no centro, incluindo o que está em serviço
 - Número de clientes esperando serviço $Q_k - U_k$
- Q : número médio no sistema

$$Q = \sum_k Q_k$$

ou usando Lei de Little (próxima aula)

Modelos com Múltiplas Classes: Entradas

- Descrição da Carga

- Número de Classes de Clientes C
- Para cada classe c :
 - Transação: taxa de chegada λ_c
 - Batch: número de clientes/jobs N_c
 - Interativo: número de clientes N_c e think time Z_c
 - Premissa: política de escalonamento independe das classes

- Descrição dos Centros de Serviço

- Número de centros de serviço K
- Para cada centro k : o tipo (LI, LD, delay center)

Modelos com Múltiplas Classes: Entradas

- Descrição das Demandas (Carga)
 - Para cada classe c e centro de serviço k
 - Tempo de serviço $S_{c,k}$ e número de visitas $V_{c,k}$ ou
 - Demanda por serviço $D_{c,k} = S_{c,k} \times V_{c,k}$
 - Demanda total de um cliente da classe c

$$D_c = \sum_{k=1}^K D_{c,k}$$

Modelos com Múltiplas Classes: Saídas

- **Medidas do Sistema: agregadas e por classe**
 - R e R_c : Tempo médio de resposta total e para classe c
 - X e X_c : Throughput do sistema total e para classe c
 - Q e Q_c : Número médio de clientes no sistema total e para classe c
- **Medidas por Centro: agregadas e por classe**
 - U_k e $U_{c,k}$: utilização do centro k agregado e para classe c
 - R_k e $R_{c,k}$: tempo médio de residência no centro k para todos clientes e para clientes da classe c
 - X_k e $X_{c,k}$: throughput do centro k para todos clientes e para clientes da classe c
 - Q_k e $Q_{c,k}$: tamanho médio da fila no centro k total e para clientes da classe c

Exercicios

Apresente o modelo (pictorico) de filas para cada problema abaixo:

- 1) Suponha um servidor de banco de dados com 1 CPU e 1 disco. Transacoes SQL chegam ao servidor para execucao a uma certa taxa λ transacoes/seg. Durante execucao, uma transacao alterna entre CPU e disco. Em qualquer instante, uma transacao pode estar usando a CPU e outra o disco, enquanto outras esperam para usar um dos dois dispositivos. Sabendo que o tempo de servico na CPU e no disco sao S_{CPU} e S_{disco} , respectivamente, e que cada transacao realiza 8 operacoes de I/O, em media, estime o tempo medio de resposta para uma transacao tipica.
- 2) Suponha o mesmo servidor de banco de dados do problema 1. Porem, tudo o que voce sabe sobre as demandas por recursos de uma transacao tipica é o tempo medio de execucao T .

Exercicios

- 3) Considere que o servidor de banco de dados do problema 1. Suponha que as transacoes SQL impostas pelos usuarios executam ao mesmo tempo que uma aplicacao batch para geracao de relatorios. O servidor sera capaz de suportar ambas as cargas, atendendo os SLA's pre-estabelecidos para cada uma delas?
- 4) Considere que o servidor do problema 1 seja agora usado como componente de uma aplicacao cliente/servidor. Maquinas dos clientes estao conectadas ao servidor de banco de dados por uma LAN. Clientes trabalham de forma independentes, alternando entre processamento local (criando requisicoes para submeter ao servidor) e esperando por uma resposta. Determine o tempo medio de resposta do servidor desta aplicacao.

Exercícios

- 5) Considere que o servidor de banco de dados do problema 1 é multithreaded. Cada thread trata uma única transação por vez e há um número limite de threads que podem executar em paralelo. Se todas as threads estiverem ocupadas, transações que chegarem deverão esperar até que uma thread fique disponível. Determine o tempo médio de resposta deste servidor se o limite de threads é igual a m . Note que seu modelo deve capturar tanto a contenção de software (espera por threads) quanto a contenção em hardware (espera por recursos).