Introdução a Avaliação de Desempenho

Avaliar é pronunciar-se sobre as características de um certo sistema. Dado um sistema real qualquer, uma avaliação deste sistema pode ser caracterizada por toda e qualquer observação sobre ele expressada. Pode-se distinguir dois tipos básicos de avaliações:

- avaliação qualitativa: neste tipo de abordagem existe a necessidade de uma comparação com o senso-comum, ou ainda uma comparação com um referencial de base.
- avaliação quantitativa: baseia-se na formulação de valores específicos, sem expressar considerações dos méritos dos valores obtidos.

Exemplo:

Suponha que o alvo de avaliação seja um programa de computador e que o aspecto escolhido para análise seja a sua modularidade. Suponha ainda que dois grupos diferentes realizem a avaliação e cheguem cada um as suas conclusões, que são as seguintes:

Grupo 1: o programa contém 2 módulos (avaliação quantitativa). Isto expressa somente uma característica do sistema avaliado, sem emitir opiniões sobre o fato.

Grupo 2: o programa não é modular pois contém poucos módulos (avaliação qualitativa). A afirmação parte do princípio que o número de módulos existente é pequeno, mas para chegar a tal conclusão é necessária uma comparação com um referencial base (por exemplo, supondo que o número mínimo aceitável de módulos seja cinco). No presente caso foi emitida uma opinião sobre o sistema avaliado.

Objetivos da avaliação de Sistemas

A princípio, toda avaliação tem por objetivo o estabelecimento de um julgamento qualitativo sobre o sistema avaliado. No entanto, toda avaliação científica é feita sobre resultados quantitativos e deve ser, tanto quanto possível objetiva, deixando para o usuário final da avaliação o julgamento do sistema avaliado.

A aplicação prática da avaliação de desempenho é o conhecimento da situação (estado) do sistema avaliado. Tanto situações anteriores como situações atuais podem ser avaliadas para tornar possível a observação da evolução do sistema. Além disso, a observação do comportamento do sistema ajuda a entender o funcionamento do mesmo. Podem ser ainda avaliadas situações futuras, com a finalidade de previsão e planejamento.

Ainda dentro do contexto de avaliação de sistemas, cabe salientar que é sempre recomendável um estudo da confiabilidade do método; para este fim é freqüente realizar-se a comparação de resultados de diversos métodos diferentes.

Interpretação dos resultados da avaliação

Tão importante quanto a avaliação é a interpretação dos resultados obtidos. Os resultados são eminentementes quantitativos enquanto que o objetivo da avaliação tem caráter qualitativo.

Em geral, bem mais importante do que o valor absoluto de um parâmetro é o seu comportamento de acordo com as variações do sistema, ou seja, a sua variação segundo alterações no modelo (ou a sensibilidade aos dados de entrada, por exemplo).

Métodos de avaliação de desempenho

Basicamente existem2 tipos de métodos de avaliação de desempenho: os métodos *elementares* e o métodos *indiretos*.

Métodos Elementares

Avaliam diretamente a realidade através de instrumentos físicos, por exemplo, com maior ou menor grau de refinamento.

Principal desvantagem: não podem ser aplicados em previsões, pois necessitam da realidade, ou pelo menos um protótipo para avaliar . Além disso, às vezes a medição direta da realidade, apesar de ser a maneira mais simples de avaliar, pode ser muito complicada ou mesmo impossível. Exemplos: medir a temperatura no interior de um reator nuclear, medir a velocidade dos ventos no interior de um tornado, etc.

Métodos Indiretos

Avaliam uma descrição da realidade, um modelo. O método de avaliação é aplicado sobre o modelo e todos os resultados obtidos serão função deste modelo.

Principal desvantagem: a falta de precisão que está ligada a construção do modelo, pois a qualidade destes métodos depende da qualidade do modelo desenvolvido e da qualidade da medição.

Um modelo não representa completamente a realidade. O “gap” semântico sempre permanece (distância entre o significado real e o significado da representação do real).

O processo de modelagem baseia-se na abstração. Esta se dá em duas etapas: primeiro são identificadas as características mais importantes (para aquele que modela) da realidade em questão e em seguida (segunda etapa) é feito o mapeamento desta realidade para o modelo que irá ser avaliado.

Os métodos indiretos dividem-se em:

Simulação: são semelhantes aos métodos elementares, pois a avaliação baseia-se na observação do funcionamento do modelo. A grande vantagem da simulação é a nível de facilidade na medição por exemplo, embora o modelo não considere a totalidade dos aspectos da realidade. Este método indireto possui normalmente um baixo nível de abstração.

Métodos analíticos: a partir de um modelo definido segundo algumas hipóteses de funcionamento, um conjunto de equações é obtido. Tais equações são a expressão matemática do modelo. Os métodos analíticos possuem o mais alto nível de abstração. Sua principal desvantagem deve-se ao fato de que geralmente suas hipóteses de funcionamento costumam são restritivas demais e a elaboração do modelo tende a ser mais complexa do que em relação ao método anterior (simulação).

Cada método possui suas características próprias. Evidentemente para cada caso real a analisar, os diversos métodos serão mais ou menos adequados. É possível estabelecer uma comparação genérica entre os métodos:

Método	Objeto avaliado	Nível de abstração	Velocidade de avaliação	Fator de dependência	Precisão dos resultados
Elementar	realidade	nenhum	real	Tempo de observação	Real
Simulação	Modelo funcional	baixo	baixa	Tempo de simulação	Alta
Analítico	Modelo comportamental	alto	alta	Complexidade algorítmica	Exata

Introdução a Teoria das Filas

Um sistema de filas (queueing system) consiste de um ou mais servidores que fornecem um tipo de serviço para clientes. Clientes que chegam no sistema e encontram todos servidores ocupados podem geralmente entrar em uma ou mais filas (ou linhas) , daí o nome de sistema de filas.

Historicamente, uma grande porção de todos estudos de simulações discretas orientadas a eventos desenvolvidos até hoje envolveu a modelagem de sistemas de filas do mundo real, ou então pelo menos um componente do sistema simulado era um sistema de filas.

Características dos Sistemas de Filas

Os elementos chave de um sistema de filas são os ***clientes*** e os ***servidores***. O termo ***clientes*** pode se referir a pessoas, partes, máquinas, aviões, processos de computador, entre outros. ***Servidores*** são caixas de banco, operadores de máquinas, controladores de tráfego, operadores de computador, etc. Outros termos importantes são:

- População: conjunto potencial de clientes; pode ser finito ou infinito.
- Capacidade do Sistema: o limite do número de clientes que o sistema pode acomodar em um dado instante de tempo.
- Processo de chegada: as chegadas podem ocorrer em tempos programados ou em tempos aleatórios, sendo que no segundo caso normalmente assume-se alguma distribuição de probabilidade. A distribuição Poisson é a mais comum.
- Disciplina de fila: o comportamento da fila em reação ao seu estado atual ou a maneira como a fila é organizada pelo servidor.
- Mecanismo de serviço (atendimento): o tempo de atendimento (service time) pode ser constante ou ter uma duração randômica . O atendimento pode se dar através de um só canal ou através de múltiplos canais.

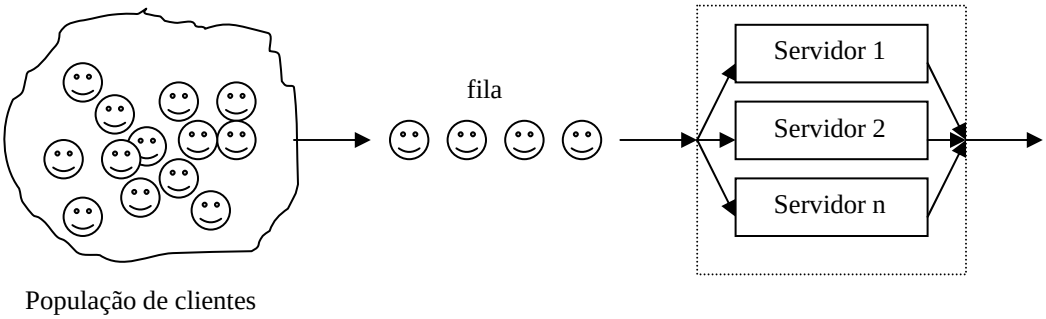
Observações importantes:

Uma variável importante do processo de chegada é a taxa de chegada (λ) dos clientes no sistema de filas. Esta taxa especifica que, por exemplo, 10 clientes por segundo vão chegar no sistema (e possivelmente serem atendidos ou então podem entrar em uma ou mais filas).

As disciplinas de filas se referem as regras que o servidor vai empregar para decidir qual será o próximo cliente da fila a ser atendido. As disciplinas mais comuns são:

FIFO: First-In, First-Out; também chamada FCFS (First-come-First-served)
LIFO: Last-In, First-Out; comportamento de pilha

Layout típico de um sistema de filas:



Notação standard para sistemas de filas (Kendall Notation)

A/B/c/K/m/Z onde:

A: distribuição do tempo entre-chegadas (interarrival time)
B: distribuição do tempo de atendimento (service time)
c: número de servidores (paralelos)
K: capacidade das filas
m: número de clientes na fonte (tamanho da população)
Z: disciplina da fila

Por exemplo, um sistema de filas
M/M/1/∞/∞ indica um sistema com :

- tempo entre-chegadas distribuído exponencialmente
- tempo de atendimento também distribuído exponencialmente
- 1 servidor
- fila de capacidade ilimitada
- população de tamanho infinito

Este tipo de sistema é usualmente abreviado como **M/M/1** , e a ausência das duas últimas letras (K , m) indicam o uso dos valores típicos (default) para infinito.

Observações:

1) o M é usado para denotar distribuição exponencial por causa de sua propriedade Markoviana de não ser influenciado por estados anteriores (memoryless)
2) o Z é considerado FIFO por default (FCFS), não sendo obrigatória sua presença neste caso.

Medidas de desempenho de sistemas de filas

As seguintes variáveis são empregadas na definição e cálculo de desempenho de um sistema de filas:

C : número de servidores do sistema
λ : taxa média de chegada de clientes
μ : taxa média de atendimento (serviço) por servidor
a : número de servidores necessários para o serviço
ρ : taxa de utilização do servidor; é uma medida de congestionamento do servidor
se $\rho < 1$ então não há congestionamento
se $\rho = 1$ então sistema está em equilíbrio
se $\rho > 1$ então há congestionamento

Wq : descreve o tempo gasto por um cliente na fila
Ws : descreve o tempo gasto por um cliente durante atendimento (serviço)
W : descreve o tempo total de um cliente no servidor (fila + atendimento)

Lq : descreve o número de clientes na fila
Ls : descreve o número de clientes em atendimento (serviço)
L : descreve o número total de clientes

Fórmulas genéricas:

$a = \lambda.Ws$
$\rho = a/c = \lambda/c.\mu$
$W = Wq + Ws$
$L = Lq + Ls$
$L = \lambda.W$
$Lq = \lambda.Wq$
$Ls = \lambda.Ws$

Para o caso M/M/1, as fórmulas específicas ficam:

$\rho = \lambda.Ws$
$W = Ws / (1-\rho)$
$Wq = \rho.W$
$L = \lambda.W = \rho/(1-\rho)$
$Lq = \lambda.Wq = \rho^2 / (1-\rho)$

$P[L = n] = (1-\rho)\ \rho^n$ probabilidade do sistema ter n clientes
$P[L \geq n] = \rho^n$ probabilidade do sistema ter n ou + clientes

Exemplo:

Suponhamos um pedágio onde há somente uma caixa fazendo o atendimento; os carros chegam a uma taxa de 2 carros por minuto e o tempo médio de atendimento de cada carro por parte da caixa é de 10 segundos.

Logo temos:

c = **1**	número de servidores (só há uma caixa, logo c=1)
λ = 2 / 60 = **0,0333**	2 carros a cada 60 segundos
Ws = **10** segundos	(tempo de atendimento ou serviço)
μ = 1/10 = **0,1**	é a taxa média de serviço; a cada segundo são atendidos 0,1 carros

a = 0,0333 . 10 = **0,333** servidores são necessários para o atendimento
ρ = 0,0333 . 10 = **0,333** taxa de utilização = 33,3 % (há sub-utilização do servidor)

Este é um caso M/M/1, portanto:

W = 10 / 1 **-** 0,333 = 10 / 0,66667 = **14,99** $\cong$ 15 segundos (é o tempo de permanência de cada cliente no sistema $\Rightarrow$ fila + atendimento)

Wq = 0,333 . 14,99 = **4,99** $\cong$ 5 seg (é o tempo médio de permanência na fila)

L = 0,0333 . 14,99 = **0,499** $\cong$ 0,5 clientes (é o número médio de clientes no sistema)

Lq = 0,0333 . 4,99 = **0,16** clientes (é o número médio de clientes na fila)

Ls = 0,0333 . 10 = **0,333** clientes (é o número médio de clientes sendo atendidos)

Qual a probabilidade da fila ter 0,5 clientes ?

$P[L = 1,5] = (1 - 0,333) . 0,333^{1,5}$ = **0,128**

E qual a chance de ter 0,5 ou mais clientes ?

$P[L \geq 1,5] = 0,333^{1,5}$ = **0,192**

Sistema em Equilíbrio (Steady State)

Normalmente nos preocupamos com a operação de sistemas de filas cujo funcionamento está “estabilizado” . Neste momento dizemos que o sistema atingiu o equilíbrio (steady state).

Quando um sistema de filas é inicializado, ele passa por um período inicial de operação que geralmente não reflete seu comportamento normal (típico). Por exemplo, quando queremos dirigir um carro por uma auto-estrada a 100 Km/h, precisamos primeiro acelerar o carro até esta velocidade. Este período inicial de operação não é representativo de como o carro vai se comportar quando estiver a 100 Km/h, pois neste período inicial o consumo de combustível é mais alto, a troca de marchas é mais freqüente e uma atenção maior é exigida do motorista. Assim que o carro atingir a velocidade desejada e esta se estabilizar, o consumo tende a diminuir e o motorista pode relaxar um pouco (mas não demais). A partir deste instante diz-se que o sistema entrou em equilíbrio (steady state).

Da mesma forma, sistemas de filas passam por um período inicial (período transiente) antes de entrarem em um contexto de funcionamento estável e previsível. A solução e o estudo de sistemas de filas é muito mais simples quando sabemos que o sistema está em equilíbrio. Neste contexto, a maioria dos parâmetros importantes se mantêm fixa, o que torna a análise do sistema bem mais fácil.