**Métricas para avaliação de desempenho**

**Caio C. Fernandes¹, Felipe C. Sacramento¹, Felipe R. V. Gouveia¹, Nagib M. C. Neto¹, Otávio A. A. Silva¹**

**¹Faculdade de Engenharia da Computação e Telecomunicações - Instituto de Tecnologia - Universidade Federal do Pará**

**Cidade Universitária “Prof. José da Silveira Netto” – Campus II (Profissional) Belém/PA, Brasil, 66075-110**

{caiocesarufpa,felipeccastro94,felipegouveia3,nmatni,tavioalves}@gmail.com

***Abstract.*** *This meta-paper will briefly describe processes and metrics for performance evaluation of computer networks. With emphasis on Markov chain, queuing theory and performance metrics for ftp, video, www (world wide web) and voice.*

***Resumo.*** *Este meta-artigo descrevera breves processos e métricas para avaliação de desempenho de redes de computadores. Dando ênfase em processos markovianos, teorias de filas e métricas de desempenho para ftp,vídeo, www (world wide web) e voz.*

1. Cadeia de Markov - Informações Gerais

A cadeia de Markov é um processo estocástico com a propriedade de Markov em um espaço de estados finito. O termo "cadeia de Markov" refere-se à sequência de um processo que move-se através de estados. Normalmente, uma cadeia de Markov é definido por um conjunto discreto de vezes.

1.1 Processos Estocátiscos

O Processo Estocatico é uma coleção de variaveis randômicas e é comumente utilizada no desenvolvimente de um sistema sobre o tempo. Basicamente Processos Estocásticos podem ser classificados de duas formas:

* Em relação ao Estado
* Em relação ao Tempo

1.2 Processos Markovianos

Segundo Nogueira, F (2009) um processo é considerado Markoviano se e somente se o processe estocastico obdecer a seguinte regra:

P{X( t k +1 ) ≤ x k +1 X( t k ) = x k , X( t k −1 ) = x k −1 ,..., X( t 1 ) = x 1 , X( t 0 ) = x 0 } = P{X( t k +1 ) ≤ x k +1 X( t k ) = x k } (1)

para t 0 ≤ t 1 ≤ ...t k ≤ t k +1 = 0,1,...

e toda seqüência

k 0 , k 1 ,..., k t −1 , k t , k t +1

Essa expressão significa que a probabilidade de um evento futuro é independente de um evento passado, porém é dependente de um evento no estado presente. Também conhecido como processo sem memória (*memoryless process*), o estado passado não vai influenciar então deve ser desprezados.“O estudo de Processos Markovianos é de extrema relevância em diversas áreas de estudo para quantificar eventos que possuem certo grau de incerteza de ocorrência” (Zimmermann,F. J. P., 2006)

1.2.1 Cadeia de Markov

Um Processo Markoviano é uma Cadeia de Markov quando as variáveis randômicas estão no estado discreto.

P{X(k + 1) = x k +1 X(k ) = x k , X(k − 1) = x k −1 ,..., X(1) = x 1 , X(0) = x 0 } = P{X (k + 1) = x k +1 X (k) = x k } ∀ seqüência 0,1,..., k − 1, k, k + 1 (2)

“As Probabilidades de Transição P{X(k + 1) = x k +1 X(k ) = x k } representam, portanto, a probabilidade do estado X(k + 1) ser x k +1 no tempo k + 1 dado que o estado X(k ) é x k no tempo k.” ( Nogueira, F., 2009)

Para simplificar a notação foi adotado  ou  de j e  de i, pode-se definir:

 (3)

e

 (4)

Estas precisam ser positivas e precisam satisfazer as seguintes propriedades:

∀  (5)

e

∀  (6)

Pode ser feita através de uma tabela P(n):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estado | 0 | 1 | … | M |
| 0 |  |  | ... |  |
| 1 |  |  | ... |  |
| ... | ... | ... | ... | ... |
| M |  |  | ... |  |

**2. Teoria de filas**

**2.1 Conceituação e Caracterização das filas de espera**

A teoria de filas vem oferecer análises matemáticas e provê modelos que fazem a análise prévia do comportamento de um sistema para uma demanda crescente aleatória, possibilitando melhorar o serviço e evitar desperdícios e gargalos.

Na fila, existe o cliente e o canal de atendimento. O primeiro é aquele que solicita o atendimento e, quando temos uma demanda maior do que a capacidade de atendimento, teremos a formação da fila de espera. Sobre esses Clientes da fila, é necessário conhecer a distribuição de probabilidade que descreve os tempos entre as chegadas dos clientes, essa que pode ser exponencial (Poisson) ou simultânea (em batch). Esse cliente então, pode optar por esperar na fila ou mesmo mudar de servidor (caso essa opção seja suportada pelo sistema).

O segundo, o canal de atendimento, ou mesmo atendente, executa as requisições do Cliente (podendo ser, por exemplo, um serviço de impressão na rede). Entre o início e o fim do atendimento, existe também uma distribuição de probabilidade sobre o tempo de serviço, esse que pode ser independente ou dependente da quantidade de clientes na fila, ou seja, a execução da requisição pode ser mais rápida ou mais lenta no servidor de acordo com a demanda. Para agilizar o atendimento, pode ser utilizado mais de um servidor para servir paralelamente dois ou mais clientes. Assim, pode haver uma fila e vários atendentes (como numa fila de banco) ou várias filas e vários atendentes (como num supermercado). Quando todos os clientes são a atendimentos imediatamente, temos o caso de um “Centro de Atraso” ou “Serviço infinito”.

No caso da atendimento, ainda temos que avaliar a disciplina no atendimento, ou seja, a forma que acontece a saída dos clientes da fila de espera para serem atendidos, estas que podem ser:

* FCFS (*First Come, FirstServed*): Primeiro a Chegar, Primeiro a ser Atendido. Disciplina mais comum, inclusive na vida diária.
* [FIFO](http://pt.wikipedia.org/wiki/FIFO) (*First In, First Out*): Primeiro a Entrar, Primeiro a Sair).
* [LCFS](http://pt.wikipedia.org/wiki/LCFS) (*Last Come, FirstServed*): Último a chegar, Primeiro a ser Atendido
* [LIFO](http://pt.wikipedia.org/wiki/LIFO) (*Last In, First Out*): Último a Chegar, Primeiro a Sair. Aplicável em sistemas em que o item mais recente é mais fácil de ser recuperado, como por exemplo em sistemas de controle de estoque.
* Fila com prioridade: a cada cliente é atribuída uma prioridade; clientes com maior prioridade têm preferência no atendimento. Pode ser de dois tipos:
  + [Preemptivo](http://pt.wikipedia.org/wiki/Preemptividade): o cliente com maior prioridade é atendido imediatamente, interrompendo o atendimento ao cliente com menor prioridade. Ao terminar, o cliente de menor prioridade volta a ser atendido, podendo continuar o processo de onde parou ou então reiniciá-lo
  + Não-preemptivo: o cliente com maior prioridade é colocado no início da fila, recebendo o serviço somente quando o cliente em atendimento sai do sistema, mesmo se este for de prioridade mais baixa
* [Round-robin (algoritmo)](http://pt.wikipedia.org/wiki/Round-robin_(algoritmo)): cada cliente recebe uma fatia de tempo do servidor (*quantum*), dentro da qual é atendido. Após o término do *quantum*, se a atividade não foi completada, o cliente é retirado e outro passa a ser atendido. Posteriormente, o cliente que foi interrompido retorna ao servidor e continua a sua atividade. É muito comum em [escalonamento de processos](http://pt.wikipedia.org/wiki/Escalonamento_de_processos) da CPU.

**3. Métricas de desempenho**

**3.1. Protocolo FTP**

**3.1.1. Servidores FTP**

O acesso a servidores FTP podem ocorrer de duas maneiras: através de uma interface ou através da linha de comando. Cabe ao usuário do serviço escolher qual forma é eficaz na aplicação, visto que, tanto usuários UNIX como usuários Windows podem acessar através dos dois modos. O modo linha de comando está presente em qualquer distribuição UNIX e Windows, através do telnet.

A partir de qualquer browser que suporte a aplicação (Internet Explorer, Firefox, ou mesmo no Windows Explorer), também é possível manter conexão a um servidor FTP, basta digitar na barra de endereço:

ftp://[username]:[password]@[servidor]

A partir de uma conexão FTP estabelecida, dois canais de transmissão serão abertos:

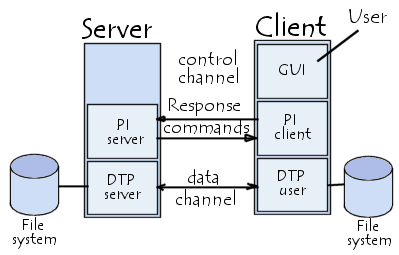
* Um canal para os comandos (canal de controle)
* Um canal para os dados

Figura 3.1.1 - Representação de FTP

Dessa forma, tanto o cliente como o servidor, possuem dois processos que permitem suprir estes dois tipos de informações:

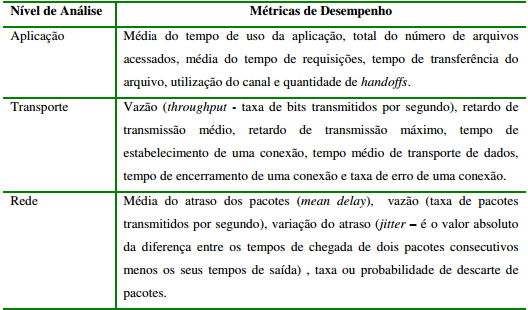
O**DTP** (*Data TransferProcess*) - É encarregado de estabelecer a conexão e gerir o canal de dados. O DTP do lado do servidor chama-se SERVER-DTP, o DTP lado cliente é denominado USER-DTP

O **PI** (*ProtocolInterpreter*) -Permite comandar o DTP com a ajuda de comandos recebidos no canal de controlo. É diferente no cliente e no servidor:

* O SERVER-PI –É encarregado de “ouvir” os comandos que chegam de uma PI no canal de controle numa porta data,
* O USER-PI – É encarregado de estabelecer a conexão com o servidor FTP, enviar os comandos FTP, receber as respostas do SERVER-PI e de controlar USER-DT, caso faça-se necessário.

**3.1.2. Métrica de desempenho FTP**

A definição das métricas a serem utilizadas depende dos objetivos da avaliação de desempenho e da carga da aplicação. Por um lado, o objetivo pode ser exclusivamente avaliar o desempenho da rede estabelecida. Sendo assim, deve-se definir em que nível se pretende medir o desempenho. Além destas, outras métricas, no entanto, são destinadas a medir o desempenho dos protocolos de acesso ao meio. Na tabela abaixo apresentamos algumas métricas usadas:

Figura 3.1.2 - Conceituação dos níveis de análise

**3.2. Vídeo**

Os sinais de vídeo são compostos de uma combinação de pixels (pontos na tela que em conjunto formam a imagem), dependendo da complexidade, o que inclui os recursos de cores, tem-se um maior ou menor número de pixels por imagem,aumentando o grau de dificuldade para os sistemas computacionais armazenarem e processarem as imagens.

Cada pixel pode ter:

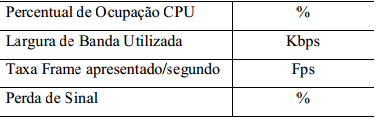
* Um único bit por pixel, para representar um ponto preto ou branco;
* 8 bits por pixel, para representar 256 tons de cinza;
* 8 bits para cada uma das cores RGB (Red=vermelho, Green=verde,Blue=azul)
* Ou 24 bits por pixel para imagens coloridas.

Em um processo normal de observação de imagens, o ser humano interpreta o movimento através de uma seqüência de quadros, os quais se formam na retina e são interpretados pelo cérebro. Para se ter a noção do movimento, são necessários a observação de pelo menos 50 quadros por segundo (fps), comparativamente sabe-se que uma média razoável na transmissão de sinais equivale a 30 quadros por segundo.

Portanto, quando se faz referencia a imagem de vídeo, a primeira idéia que surge é da tela de uma televisão, com tamanho normal de média igual a 20 polegadas, no intentonas considerações sobre sinais de vídeo em redes de computadores está se considerando uma parcela menor da tela que normalmente é utilizado por um software para a reprodução das imagens.

**3.2.1 Métricas de desempenho de Vídeo**

As métricas de desempenho de vídeo são, relativamente simples de serem mensuradas, visto que estas são mais ligados a fatores externos. Portanto, é fundamental apontar variáveis como: largura de banda; algoritmo de codificação e número de participantes em uma sessão. Entretanto, quatro fatores essenciais devem ser mensurados também para a real medida de desempenho:

Figura 3.2.1 - Medidas de desempenho

**Percentual de Ocupação da CPU** – tempo que os diversos microcomputadores, individualmente, levam para processar as imagens de vídeo que estão sendo recebidas.É levado em consideração o tempo do processo de descompactação e apresentação na tela, desprezando, no entanto,a memória utilizada nestas atividades.

**Largura de Banda Utilizada** - Considera-se o fato de ter o alto nível de exigência das aplicações de vídeo,desta forma é esperado o valor em Kbps que represente areal ocupação e de indícios sobre o potencial de utilização do canal de rede.

**Taxa Frame apresentado por segundo** – O valor para esta métrica apresenta o resultado real que uma aplicação de vídeo irá resultar.

**3. 3. Métricas de desempenho - WWW(World Wide Web)**

Com o uso excessivo desse serviço por vários setores, se tornou notável a preocupação em se melhorar e adequar o uso e a necessidade dessa plataforma, já que uma súbita queda dessa ferramenta pode ser crucial para os lucros de uma empresa, a qualidade de um serviço ou perda de dados em questão.Portanto logo foram criadas meios para se medir a qualidade do serviço, conhecidas como métricas. Para se realizar este estudo, o engenheiro responsável deverá prestar atenção em o seu desempenho, o seu comportamento em certas situações corriqueiras, a sua capacidade de fluxo de dados e a onde esse tipo de plataforma deve ser usada para solucionar um problema.

**3. 3. 1. Requisitos para realização da métrica**

Para se realizar as métricas de desempenho, que devem se ter primeiramente os requisitos que serão utilizados nas metas, como por exemplo:

* Total de Visitas por semana
* Total Hits por semana
* Total de sessões de usuários por semana
* Tamanho da página Média
* Tamanho Hit média
* Distribuição de solicitação da página
* Saída do usuário da pagina

Depois de se possuir os requisitos das métricas é necessário saber como o servidor irá se comportar com o usuário desse serviço por tanto são utilizados alguns fatores para os testes, como por exemplo:

* Velocidade interação
* Tolerância de latência
* Familiaridade
* Velocidade de conexão
* Localização

**3.3.2. Aplicação da métrica de desempenho - WWW**

Depois da coleta de informações a respeito dos requisitos e informações do usuário, podemos passar para as métricas de desempenho , abaixo está algumas métricas existentes com os seus respectivos conceitos.

**Monitoramento do VUSER**: A partir dessa métrica podemos saber  estado de cada usuário em cada operação executada para analisar o impacto da crescente carga de trabalho no sistema.

**Monitoramento de transações**: As transações são processos de negócios que são executadas para determinado usuário para testar o desempenho do aplicativo. É vital saber o estado de cada operação executada e seu tempo de resposta.

**Recursos Web**: Métricas de recursos da Web fornecem o servidor web informações de parâmetros capturados durante o teste de desempenho.

**Monitoramento de recursos do sistema**: Acompanhamento de todas as máquinas (Servidores e injetores de carga) e os seus recursos de sistema operacional que estão sendo usados ​​em testes de desempenho é também fornece informações úteis na análise.

**Métricas de desempenho do Servidor web** : Métricas de servidor Web fornece as informações de uso de recursos em servidor web durante o teste de desempenho. Essas métricas fornece a informação útil sobre o desempenho do servidor web e seus gargalos.

**Aplicação métricas de desempenho do servidor** : A maior parte das atividades de aplicações computacionais são realizados no seu servidor de aplicação. Podemos dizer servidor de aplicações é o osso de volta da aplicação, especialmente no caso de aplicações de negócios de alta complexidade. Alguns fatores importantes para as métricas do desempenho do servidor são: Memória livre e total, uso de memória, tópicos ativos, total de tópicos, tempo da conexão e sua espera, percentual de conexão usada, transações ativas e suspensas, timeouts, sessões invalidas e etc.

**3.4. Métricas de desempenho em Voz**

Será utilizado neste artigo o VoIP, como exemplo da análise de desempenho em um sistema de voz. Com o intuito de analisar o desempenho de um sistema VoIP (Voice over Internet Protocol) para analisar a qualidade da voz recebida deve-se analisar as seguintes métricas:

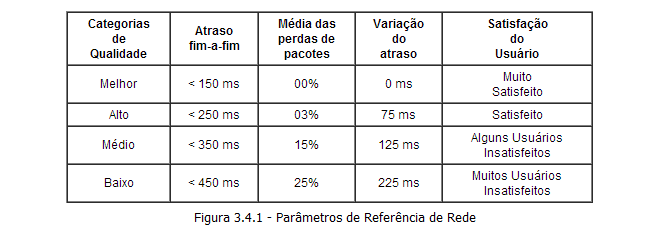
**1) Vazão:** taxa de transmissão de bits por segundo;

**2) Atraso por pacote:** tempo que o pacote leva da origem até o destino;

**3) Variação no atraso por pacote:** variação do atraso por pacote subjacente;

**4) Taxa de perda de pacotes:** quantidade de pacotes que foram perdidos durante a transmissão de dados;

A tabela da figura 6.1 relaciona os valores limites e os níveis de qualidade de voz com base **apenas** nos parâmetros de desempenho em nível de rede. **Note**: Outros componentes como o a qualidade do dispositivo final da transmissão de Voz irão influenciar no desempenho total, nesta tabela estão relacionados apenas os quesitos de rede.



**3.4.1. Modelos Computacionais**

A qualidade de voz obtida na VoIP pode ser calculada e tabelada utilizando modelos computacionais que avaliam a qualidade conversacional na telefonia. O ***E-Model*** e a **MOS** (*Mean Opinion Score*) são exemplos destes modelos computacionais.

**3.4.1.1 – E-Model**

O *E-Model* calcula os efeitos de variações em diversos parâmetros de transmissão e resulta em um valor chamado “**Fator R**” que determinará a qualidade conversacional na telefonia, variando de 100 até 0. O *E-Model* é calculado utilizando a fórmula:



Onde:

**- Ro:** é a relação Sinal-Ruído;

**- Is:** perturbações simultâneas na transmissão;

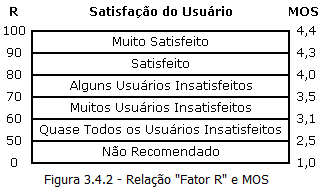
**- Id:** perturbações após a transmissão;

**- Ie:** perturbações causadas pelos codecs;

**- A:** Fator de Expectativa;

**3.4.1.2. *Mean Opinion Score* (MOS)**

Sendo o modelo mais aceito para determinar a qualidade da chamada de voz o MOS é calculado baseado em opiniões em testes realizados com grupos de pessoas. Cada indivíduo do grupo irá escutar o a*udio* e fornecer uma nota de 1 à 5, uma média aritmética é então feita, resultando no MOS. A figura 3.4.1 relaciona ambos o critério *E-Model* com o MOS.



**4. Referencias**

A. A. Markov, An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains, trans. David Link. Science in Context 19.4 (2006): 591–600.

G.Bolch, S.Greiner, H.de Meer and K.S.Trivedi, *Queueing Networks and Markov Chains*, John Wiley, 2nd edition, 2006

A.C.B.Alves, M.A.F.Menezes, F.J.P.Zimmermann, Pesquisa Operacional Um texto essencial para Engenharias, Computação e Ciências Econômicas, 2006

Tanenbaum, Andrew S.; J. Wetherall, David - Redes de Computadores - 5ª Ed.

https://repositorio.ufsc.br/bitstream/handle/

http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle

http://www.bibliotecadigital.unicamp.br/document/?code=000436385

http://www.agileload.com/agileload/blog/2013/02/18/web-applications-performance-testing-metrics

http://www.teleco.com.br/tutoriais/tutorialmondesvoip/pagina\_5.asp

http://www.teleco.com.br/tutoriais/tutorialmondesvoip/pagina\_6.asp

http://pt.wikipedia.org/wiki/Teoria\_das\_filas

http://www.sbm.org.br/docs/coloquios/CO-2.06.pdf

http://www.deinf.ufma.br/~mario/grad/filas/TeoriaFilas\_Cajado.pdf