Caracterizando variáveis de interatividade dos alunos do curso de computação do CEDERJ baseado no servidor multimídia RIO.

Bruno C. B. Alves, Rosa M. M. Leão, Edmundo de Souza e Silva

[1]Programa de Engenharia de Sistemas e Computação – COPPE/UFRJ
Caixa Postal: 68.511 – 21941-972 – Rio de Janeiro, RJ

{balves,rosa,edmundo}@land.ufrj.br

Resumo. *O rápido e expressivo aprimoramento das tecnologias de processamento e transmissão de dados tem propiciado uma popularização de aplicações como as de transmissão de objetos multimídia (vídeo, voz, etc). Estas aplicações demandam uma considerável quantidade de recursos e, mesmo com as atuais altas taxas de transmissão e recuperação de informação, é necessário o uso de técnicas de compartilhamento eficiente de recursos de forma a aumentar a escalabilidade dos sistemas, mantendo o desempenho satisfatório e conseqüentemente uma qualidade de serviço dentro de limites aceitáveis. Para que tais técnicas possam ser avaliadas e ainda dimensionar adequadamente os recursos utilizados pelas aplicações, é fundamental conhecer as características dos usuários que interagem com estes sistemas. Neste trabalho é feita uma caracterização do comportamento interativo de usuários do servidor multimídia RIO, utilizado pelo curso de Sistemas de Computação do projeto de educação superior à distância do consórcio CEDERJ de universidades públicas do Estado do Rio de Janeiro.*

1. Introdução

Os avanços científicos e tecnológicos na última década propiciaram um aumento expressivo da capacidade de transmissão e processamento de informação. Neste cenário, tornam-se cada vez mais populares aplicações como as de transmissão de objetos multimídia compostos por vídeo, áudio e imagens. Aplicações deste tipo possuem severos requisitos com relação ao retardo, jitter e a taxa de transmissão. Mesmo com as atuais altas taxas de transmissão e recuperação de informação, é necessário o uso de técnicas eficientes de compartilhamento de recursos para garantir um bom desempenho destes sistemas e aumentar a sua escalabilidade.

Para projetar adequadamente um sistema é necessária a criação de modelos que, por sua vez, exigem não somente o conhecimento dos recursos e técnicas que impactam no desempenho, mas também da carga de trabalho a que o sistema será submetido. No caso dos sistemas que estamos interessados, a carga é gerada pelos usuários deste e, portanto, é imprescindível caracterizar acuradamente o comportamento dos usuários da aplicação. A determinação dos processos que representam este comportamento e suas características estatísticas nos permitirá criar e parametrizar modelos de carga de trabalho, fundamentais para compor o modelo geral do sistema, e com isso avaliar técnicas de compartilhamento de recursos, dimensionar o sistema e permitir escalabilidade.

Este tipo de análise, apesar de essencial, é pouco encontrada na literatura, pois exige um trabalho árduo de monitoramento de um sistema em operação. Dos poucos trabalhos existentes relacionados a servidores de vídeo, uma pequena quantidade estuda criteriosamente as variáveis de interatividade dos usuários, interatividade esta natural de sistemas que permitem que um usuário tenha total controle dos objetos que estão sendo transmitidos sob demanda, em tempo real.

Destes trabalhos alguns se destacam como o de [Almeida et al. 2001], onde é feita uma análise da carga gerada por usuários dos sistemas BIBS e eTeach, usados para distribuição de conteúdo educacional em duas grandes universidades dos Estados Unidos. Ambos os servidores permitem que o aluno realize interações com o conteúdo que está sendo transmitido. Os vídeos usados são de alta qualidade e são analisadas características como popularidade, localização e freqüência dos acessos, além de características de interatividade como tempos de atividade e inatividade dos usuários nas sessões.

Um outro estudo feito em [Costa et al. 2004] analisa, além do servidor eTeach, provedores de conteúdo de entretenimento em áudio (rádio/UOL e ISP/áudio) e vídeos de curta duração (TV/UOL). O trabalho analisa características de acesso e interatividade (e.g., tempos de atividade e inatividade nas sessões, posição inicial acessada e distância dos saltos no conteúdo) dos usuários dos quatro sistemas mencionados. O trabalho de [Rocha et al. 2005] é complementar ao estudo de [Costa et al. 2004] pois é avaliado o impacto do comportamento interativo de usuários de sistemas de distribuição de conteúdo multimídia, tanto para entretenimento quanto para educação, no uso de protocolos de compartilhamento de banda. É proposto um modelo para geração de carga sintética baseado nos dados reais dos servidores de mídia de entretenimento (UOL) e de educação (eTeach e MANIC). Este modelo é usado para avaliar e propor protocolos de compartilhamento de banda para usuários interativos.

O trabalho de [Padhye and Kurose 1998] foi um dos primeiros desta natureza e avalia o sistema educacional MANIC, na sua primeira versão, que oferecia para o usuário áudio com slides sincronizados sob demanda, e permite o uso de comandos como pausar, avançar e retroceder, dentre outros. Variáveis como duração das sessões, tempos de inatividade e atividade dentro das sessões e distância dos saltos são estudadas. Já em [Tomimura et al. 2006] uma versão recente do sistema MANIC foi estudada. (O sistema MANIC é baseado em vídeo-CDs para a entrega de conteúdo, e não em servidor multimídia.) Nesta nova versão o sistema oferece, além de áudio e slides sincronizados, vídeo das aulas. Foram analisadas diversas variáveis como o tempo de inatividade e atividade dentro das sessões, o tamanho dos saltos, posição inicial acessada, tempo de permanência em slides, assim como a correlação entre algumas variáveis do sistema. Além da análise de variáveis, foi proposto um modelo baseado em cadeia de Markov oculta para geração de carga sintética.

O trabalho de [Veloso et al. 2002] se diferencia dos já citados pois a mídia é distribuída ao vivo (não é pré-armazenada e entregue sob demanda como nos sistemas dos trabalhos citados acima). Uma comparação com este estudo se torna interessante pois o sistema distribui mídia de entretenimento (”Reality TV Show brasileiro”) e o conteúdo sendo exibido é dirigido pelo sistema e não pelo usuário. Portanto, não há interatividade existente nos sistemas de vídeo/áudio sob-demanda. São estudadas características dos usuários como duração da sessão, tempos de atividade e inatividade na sessão, processo

de chegada de clientes, padrões sazonais e interesse dos clientes.

Neste trabalho são analisadas características de interatividade dos usuários do sistema multimídia RIO. Este sistema tem características distintas dos acima e está em operação desde Março de 2005. O servidor RIO permite aos usuários acesso a aulas pré-gravadas com vídeo e slides sincronizados. Os usuários são alunos do curso de Sistemas de Computação do projeto de educação superior à distância do consórcio CEDERJ (Centro de Educação Superior a Distância do Estado do Rio de Janeiro) de universidades públicas do Estado do Rio de Janeiro. Neste ambiente, os alunos têm total controle sobre a aula que estão assistindo, podendo se movimentar livremente sobre todos os tópicos de cada aula e de diversas formas, parar e retomar a exibição a qualquer instante e de qualquer ponto da aula. Foram analisadas 2674 sessões de usuários, cada sessão correspondendo a uma aula assistida por aluno, durante o ano de 2005. Como principal contribuição está o estudo detalhado e parametrização de variáveis de interatividade dos usuários do servidor RIO em um ambiente real de operação no estado do Rio de Janeiro. Apesar de a diferença do servidor em relação a outros sistemas, algumas das variáveis de interatividade são semelhantes a outros. Neste caso, uma comparação das características é também realizada. Os resultados aqui encontrados foram fundamentais no trabalho de criação de modelos de usuários e de carga de trabalho sintética utilizados na avaliação de técnicas de compartilhamento de recursos [Vielmond et al. 2007].

Inicialmente, na seção 2, o artigo apresenta uma visão geral do ambiente de ensino à distância estudado e do sistema RIO utilizado. As medidas de interesse avaliadas são descritas na seção 3 e a metodologia de análise é discutida na seção 4. Os resultados encontrados e as observações relevantes obtidas pela análise destes compõem a seção 5 e na seção 6 resumimos nossas contribuições.

2. Ambiente Estudado

Neste trabalho são analisados arquivos de log (registros de ações e acesso) dos usuários do sistema multimídia distribuído RIO, usado no curso de computação do projeto para ensino à distância CEDERJ. Este projeto, gerenciado pelo Governo do Estado do Rio de Janeiro, visa possibilitar o acesso à educação, de forma semi-presencial, de alunos de cidades do interior do estado. Foram estudados logs de acessos de alunos do curso de graduação em Tecnologia de Sistemas de Computação, elaborado em parceria entre a UFRJ (DCC/IM e PESC/COPPE) e a UFF (Instituto de Computação) de três pólos de ensino do projeto localizados na cidade de Piraí, Volta Redonda e Três Rios. Cada um dos pólos possui um servidor multimídia instalado na sua rede local.

Para assistir às aulas do curso de computação, os alunos visitam o pólo onde estão matriculados, fazem o *logon* na plataforma educacional do CEDERJ e, a partir desta, se conectam no servidor RIO para acessar uma aula de seu interesse. Cada sessão corresponde à visualização de uma aula do curso. Cada interação do aluno com a aula é gravada em um arquivo e, quando a sessão é encerrada, este arquivo é armazenado no servidor do pólo.

O sistema utilizado para armazenar, gerenciar e disponibilizar o conteúdo do curso foi desenvolvido pelo Laboratório LAND da COPPE/UFRJ, a partir de um protótipo inicial projetado em parceria com a UCLA e a UFMG. O servidor multimídia RIO (Random

I/O System) [Netto et al. 2005] possui como umas de suas características principais o armazenamento aleatório de blocos [dos Santos et al. 2000]. O RIO é composto por um servidor principal que gerencia os pedidos dos clientes e os repassa a um ou mais servidores de armazenamento que enviam diretamente ao cliente os dados solicitados, não sobrecarregando o servidor principal. O servidor principal e os de armazenamento não precisam estar localizados na mesma máquina, permitindo uma arquitetura totalmente distribuída, com os componentes em localidades distintas interligadas por rede.

Para acessar o conteúdo armazenado no servidor, os usuários utilizam um software cliente (desenvolvido pelo LAND/UFRJ) que possibilita a interação do aluno com o conteúdo que está sendo apresentado. As aulas usadas no projeto CEDERJ são compostas por vídeo e slides sincronizados e a transmissão se dá sob demanda em tempo real. Os slides contém animações para facilitar o entendimento por parte do aluno da aula sendo assistida. Através do cliente os usuários podem paralisar a exibição da aula (pressionar *pause*); saltar para outro ponto através de ações como: *fast forward*, *fast rewind*, arrastar a barra de progresso, clicar no índice de um slide; e parar (através do comando *stop*) a exibição do conteúdo a qualquer instante. Quando o usuário deseja encerrar a sessão ele clica no comando *quit*.

3. Métricas

Nosso objetivo é encontrar funções de distribuições de probabilidade cumulativa de variáveis que representem as características de interatividade dos usuários do sistema. São avaliadas as seguintes variáveis: tempo em *play*, tempo em *play* interativo, tempo em *OFF*, tamanho dos saltos para frente e para trás, posição inicial acessada e tempo de permanência em um *slide*. Outras características como a popularidade dos objetos no servidor, freqüência das interações e duração das sessões também são analisadas. (A partir destes resultados foi possível construir um modelo de usuário que permite gerar carga sintética e avaliar técnicas de compartilhamento de banda [Vielmond et al. 2007].)

O intervalo em *play* (ou em *ON*) corresponde ao tempo em que o usuário está assistindo a um vídeo até que ocorra uma interrupção gerada pelo aluno ou então a sessão seja encerrada. Foi observado um número elevado de tempos em *play* de curta duração. Acreditamos que nestes casos, onde o usuário salta e interage com o vídeo várias vezes gerando tempos em *play* de pequena duração, um assunto específico da aula está sendo procurado. Este comportamento pode ser melhor entendido através da Figura 1 (a) onde, durante a sessão, o aluno interrompe consecutivamente a exibição do vídeo através de interações como *fast forward*, *fast rewind*, salto pela barra de progresso e pelo índice.

Para uma melhor caracterização da métrica tempo em *play*, será feita uma divisão da mesma em duas categorias. Uma, denominada *play* interativo, é composta pelo conjunto dos intervalos *play* entre saltos de "curta duração", isto é aqueles com valor inferior a um dado limite (Figura 1 (a)). A outra, denominada simplesmente de *play*, consiste do restante das amostras. Obviamente é importante estabelecer um valor para tal limite. O gráfico 1(b) representa a fração de ocorrência de intervalos em *play* entre saltos para valores do limite entre 1 a 10 segundos. Através do gráfico podemos verificar que o valor de 5 segundos engloba quase a metade das amostras em questão, além de ser um tempo suficiente para um aluno identificar o conteúdo exibido, isto é, determinar que não é a parte do assunto desejado. Baseados na Figura 1(b) escolhemos então o valor de 5 segundos

para o limite e portanto os intervalos em *play* entre saltos menores que 5 segundos serão classificados como *play* interativo.

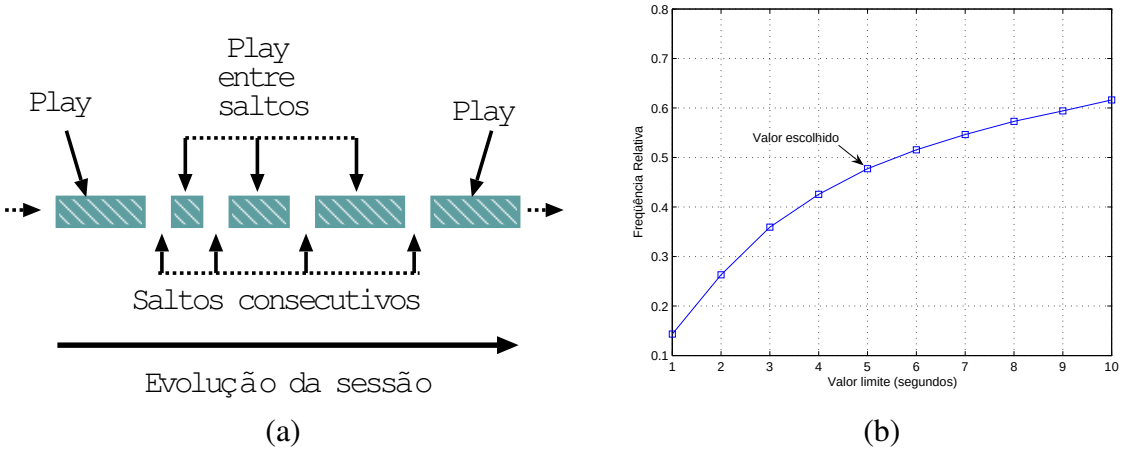


Figura 1. (a) Ocorrência de intervalos em *play* entre seqüências de saltos; (b) Fração de amostras de intervalos em *play* em função do limite.

O tempo em *OFF* se refere aos intervalos de inatividade do usuário. Esta métrica é composta pelos tempos em que a exibição é interrompida (*pause*) e parada (*stop*). A diferença entre parar o vídeo com o botão *pause* ou *stop* é que, no momento em que o botão *play* é acionado, no primeiro caso a exibição é reiniciada do mesmo ponto de parada do vídeo e, no segundo, o vídeo retorna à posição inicial da aula. Existem outras interações que interrompem o intervalo de *play*. São elas os saltos pela barra de progresso, pelo índice de tópicos da aula e pelos botões *fast forward* e *fast rewind*. Entretanto, imediatamente após cada uma destas interações o sistema continua a exibição do vídeo, não caracterizando um período de inatividade do cliente. Esta é uma das característica do servidor RIO distinta da de outros sistemas (e.g., [Tomimura et al. 2006],[Padhye and Kurose 1998]) onde a exibição do vídeo imediatamente após um comando de salto é paralisada e só é retomada quando o usuário aciona novamente a tecla *play*.

Caracterizar o tamanho dos saltos dos usuários nos possibilita entender a forma como eles se posicionam e se movimentam durante a aula. Esta característica é de extrema importância para o projeto e ajuste do sistema, uma vez que informações como a proximidade entre o destino e a origem dos saltos pode beneficiar o uso de estratégias de otimização como *cache*, *prefetching* e protocolos de compartilhamento de banda. Analisamos separadamente saltos para frente e para trás. Como já foi mencionado, os saltos podem ocorrer através do uso dos botões *fast forward* e *fast rewind*, da movimentação da barra de progresso e acionando um dentre os vários tópicos contidos no índice de conteúdo da aula. Os saltos *fast forward* e *fast rewind* possuem tamanho fixo, ou seja, no sistema estudado, ao pressionar uma única vez um desses botões, um número pré-determinado de blocos de vídeo é saltado, tanto para frente como para trás, dependendo do caso.

Quando o usuário inicia uma sessão, a única opção habilitada no cliente é a tecla *play*. A partir do momento em que *play* é acionada, o vídeo passa a ser exibido do início. Caso o aluno deseje assistir uma outra parte da aula, deve acionar uma das teclas

disponíveis, ou a barra de progresso. Devido a esta característica, isto é, uma sessão sempre é iniciada com um comando *play*, obviamente o primeiro *slide* da aula é sempre o primeiro a ser acessado. No entanto, este primeiro *slide* pode não ser o de interesse do aluno e, neste caso, haverá um salto imediato para outro ponto do vídeo. Para se ter uma estimativa da posição inicial escolhida pelos alunos, observamos qual é o **segundo** *slide* acessado. Seja esse o *slide* de número S na seqüência da aula. Caso $S = 2$ e a transição do primeiro para o segundo tenha sido seqüencial, ou seja, nenhuma interação foi realizada até o acesso ao *slide* 2, o *slide* número 1 é considerado como a primeira posição acessada pelo cliente. Caso contrário, o valor de S (correspondente ao segundo *slide* acessado pelo aluno) é considerado o inicial.

Obviamente podem haver casos em que o aluno assiste a parte do vídeo associado ao primeiro *slide* até praticamente o final e então realiza um salto para S. Pela definição acima, S seria o *slide* inicial. Mas neste caso, o aluno realmente assistiu o primeiro *slide* e então este deveria ser o *slide* inicial. Para verificar se existem casos deste tipo analisamos o valor dos intervalos até o salto, em todos os casos em que o aluno pressiona uma tecla de interação durante a visualização do primeiro *slide*. Observamos que a média foi de 3,41 segundos e mais de 80% dos casos foram menores que 5 segundos. Portanto, não estamos cometendo erros na escolha do *slide* inicial de acordo com a definição acima.

Uma outra característica analisada é o tempo que os alunos despendem assistindo um *slide* qualquer da aula. Seja T_s a duração da parte da aula enquanto o *slide* s é exibido. Um aluno pode assistir a partes do vídeo deste *slide* apenas, por exemplo saltando para um outro ponto do mesmo *slide* no vídeo. Ou retornando e assistindo várias vezes um mesmo pedaço do vídeo associado ao *slide*. Logo o valor desta métrica pode ser menor, igual ou maior que T_s, pois contabilizamos o tempo total gasto pelo usuário até que ocorra uma *troca de slide* independente se houve ou não interações com o vídeo.

4. Metodologia

O nosso objetivo principal é determinar quais as distribuições de probabilidade que caracterizam as métricas definidas na seção 3. Para isso, inicialmente filtramos os logs que não apresentavam nenhum problema. Em seguida foram estimados parâmetros de diferentes distribuições para as métricas de interesse. Finalmente, foram usados diversos métodos estatísticos para a escolha da distribuição mais adequada para caracterizar cada uma das métricas. Usamos scripts na linguagem PERL [O'Reilly and Associates], as ferramentas MATLAB [The Mathworks] e DATAPLOT [NIST] nas nossas análises.

4.1. Seleção dos logs

Primeiramente foram descartados os logs que apresentavam problemas como, por exemplo, aulas onde o índice dos slides não foi carregado corretamente, e portanto não permitindo o uso do comando de acesso através deste índice.

Após esta primeira seleção, observamos que uma fração considerável das sessões (aproximadamente 45%), tinham duração inferior a 5 minutos. Como a duração média de uma aula é de cerca de uma hora, decidimos selecionar para análise somente sessões acima de 5 minutos. Optamos por desconsiderar essas sessões muito *curtas* pois o comportamento do aluno neste caso pode ser bastante diferente do padrão dos alunos que assistem normalmente uma aula. Essas sessões de curta duração indicam que o aluno

apenas estava verificando o assunto tratado (assim como folheando um livro). Portanto não são de nosso interesse, pois o foco é o comportamento de um aluno que realmente assistiu a uma boa parte do assunto, semelhante a uma aula presencial.

Analisamos também outros valores para o filtro de duração da sessão. Selecionamos amostras de sessões com duração acima de 20, 30, 40 e 50 minutos. O uso destes outros filtros com duração superior a 5 minutos não alterou significativamente os resultados encontrados quando o filtro de 5 minutos foi usado. Portanto, os resultados das nossas análises são baseados em 1452 logs de sessões com duração acima de 5 minutos. Cabe ressaltar que analisamos os logs sem usar o filtro de duração da sessão. Neste caso a média de algumas métricas teve alterações significativas, como esperado.

4.2. Parametrização e escolha das distribuições

O primeiro passo para modelar as amostras é estimar os parâmetros das distribuições que serão testadas. O método utilizado para esta parametrização foi o *Maximum Likelihood Estimate*(MLE) [NIST/SEMATECH 2006]. Este método se baseia na máxima probabilidade (verossimilhança) da distribuição gerar valores próximos das amostras empíricas. No caso da distribuição hiperexponencial, o software EMpht [Olsson] (uma implementação de algoritmos EM interativos, que se baseiam também na máxima verossimilhança) foi utilizado. Para as outras distribuições usamos o MATLAB.

As distribuições parametrizadas foram: gamma, weibull, exponencial, lognormal, normal e hiperexponencial. A decisão a respeito do número de estágios a ser usado na hiperexponencial é feita parametrizando a distribuição para 2, 4, 6 e 8 estágios e observando os resultados de alguns testes que estão descritos a seguir.

Com as distribuições já parametrizadas, fazemos a comparação das curvas dos modelos com a curva da distribuição empírica. A primeira análise gráfica é feita usando as curvas de distribuição de probabilidade cumulativa complementar (CCDF) com os valores do eixo Y (1-F(x)) em escala logarítmica, para evidenciar a cauda das distribuições. Entretanto, a simples análise gráfica visual não deve ser considerada decisiva.

De forma a complementar a análise gráfica, estimamos o erro quadrático médio (MSE) entre cada distribuição e a curva empírica, e usamos dois testes para avaliar se uma distribuição pode ser empregada para representar uma determinada métrica.

O primeiro teste é o de hipótese de Kolmogorov-Smirnov [Trivedi 2002]. Este teste aceita a hipótese H_0 dos dados amostrais pertencerem à distribuição sendo testada, se a estatística D_{max} (que chamaremos de KSSTAT) calculada no teste for menor ou igual a um valor crítico, proveniente de uma tabela de valores predefinida para o teste. O valor de D_{max} corresponde a distância máxima encontrada entre as curvas que estão sendo avaliadas. A sensibilidade do teste é definida pelo nível de significância escolhido α, que corresponde à probabilidade de rejeitarmos a hipótese nula erroneamente. Neste trabalho usaremos α igual a 0.05. (Os valores 0.01, 0.05 e 0.10 são os mais utilizados e sugeridos na literatura).

O outro teste utilizado é o teste gráfico chamado QQPlot (Quantile-Quantile Plot) [NIST/SEMATECH 2006]. Este mostra se dois conjuntos de amostras vêm de uma mesma população, ou seja, possuem distribuições provenientes de uma mesma ”família”. Inicialmente são calculados os quantiles usando dois conjuntos de amostras: o conjunto

de amostras empíricas e o de amostras geradas segundo uma distribuição escolhida. Um quantile é a fração de amostras menores que um dado valor. No QQPlot são representados os valores dos quantiles obtidos para ambos os conjuntos de amostras. Se os dois conjuntos de amostras são provenientes de uma mesma distribuição, os pontos do gráfico devem formar, aproximadamente, uma reta com inclinação de 45 graus.

Com esses métodos, que permitem a análise visual da cauda das distribuições (gráficos da distribuição complementar), calculam a distância quadrática média entre as curvas da distribuição (MSE), computam os pontos mais distantes entre elas (teste de Kolmogorov-Smirnov), e permitem avaliar se dois conjuntos de amostras pertencem a uma mesma distribuição (QQplot), acreditamos que podemos obter resultados bastantes confiáveis. Uma análise tão criteriosa quanto essa não é observada, do nosso conhecimento, em nenhum dos trabalhos relacionados presentes na literatura.

5. Resultados

Nesta seção serão apresentadas as estatísticas e resultados numéricos obtidos da análise do comportamento dos usuários do sistema. Além das informações sobre a característica de interatividade dos usuários, apresentaremos também algumas estatísticas complementares a respeito do sistema.

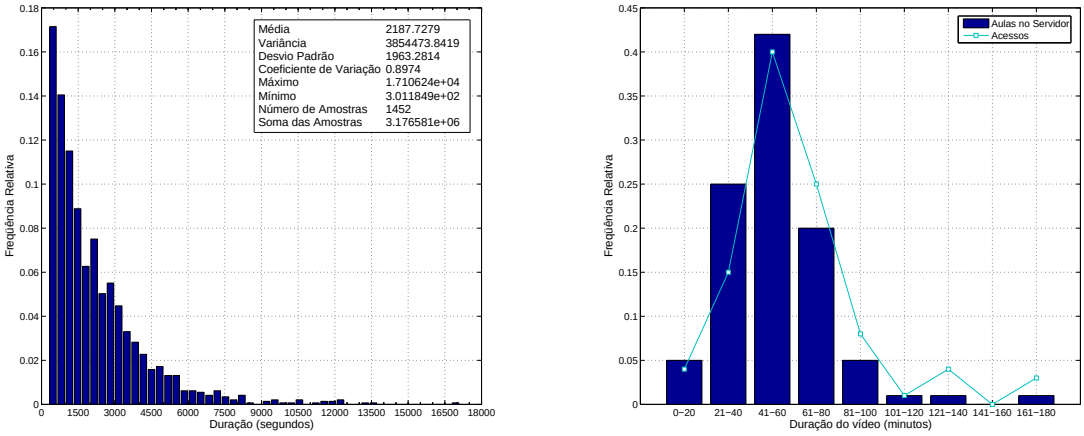
5.1. Estatísticas Iniciais

Na tabela 1 são apresentados dados do servidor RIO e algumas estatísticas que foram obtidas considerando todos os logs válidos, ou seja, não foi usado nenhum filtro com relação a duração de uma sessão.

Período de coleta	Março a Dezembro de 2005
Número de Aulas	146
Aulas armazenadas no Rio	60 Gigabytes
Número de Disciplinas	8
Número de sessões	2674
Medida	**Mínimo / Máximo / Média / Coeficiente de Variação**
Duração de uma aula	00:11:36 / 02:52:21 / 00:57:36 / 0,5
Número de slides por aula	11 / 89 / 34 / 0,36
Duração da sessão (segundos)	1,19 / 17106,23 / 1224,45 / 1,48

Tabela 1. Estatísticas básicas

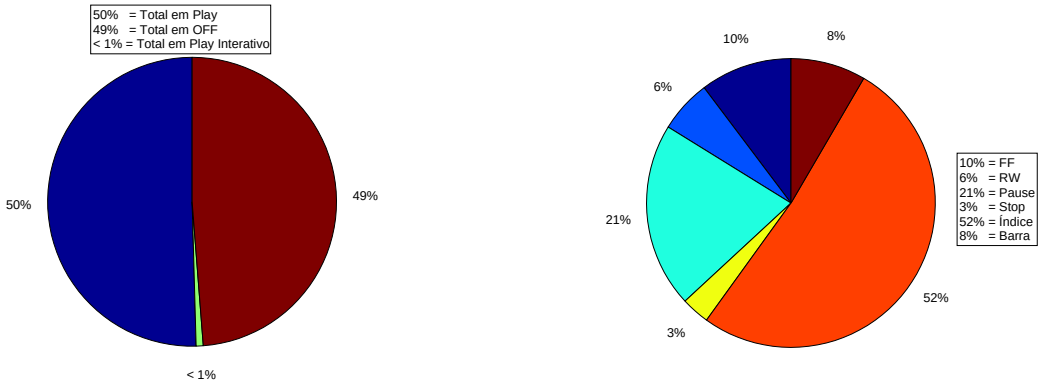
A Figura 2(a) mostra o histograma da duração das sessões considerando os logs com sessões acima de 5 minutos (Todos os resultados apresentados a seguir usam este filtro.) Podemos notar que a média da duração da sessão teve o seu valor aumentado em aproximadamente 80% quando usamos o filtro de duração da sessão (passou de 1224s para 2187s). Outras estatísticas como coeficiente de variação e valor mínimo também tiveram alterações significativas nos seus valores. Observamos também que a grande maioria das sessões são de até 1:30h. Na Figura 2(b) pode-se observar o histograma da duração das aulas armazenadas no servidor. Nota-se que mais de 40% das aulas tem duração entre 40 e 50 minutos e que essas são as aulas mais acessadas.



(a) (b)

Figura 2. (a) Histograma da duração das sessões; (b) Histograma da duração das aulas e dos acessos dos clientes.

Para ter uma idéia a respeito do comportamento interativo dos alunos, calculamos a fração de comandos executados durante uma sessão. A Figura 3(a) mostra que os alunos passam 51% do tempo ativos (com blocos de vídeos sendo tocados) e a Figura 3(b) mostra que as interações pelo índice são as mais freqüentes, seguidas pelas interações de pausa. As interações de *fast forward* e *fast rewind* costumam acontecer em seqüência, em média são realizadas 4 interações de *fast forward* consecutivas com intervalo menor do que 5 segundos entre elas. Outra característica com relação às variáveis que caracterizam a interatividade dos alunos é que a maior parte das amostras dessas variáveis apresentou coeficiente de variação maior do que 1, o que sugere o uso de uma distribuição hiperexponencial.



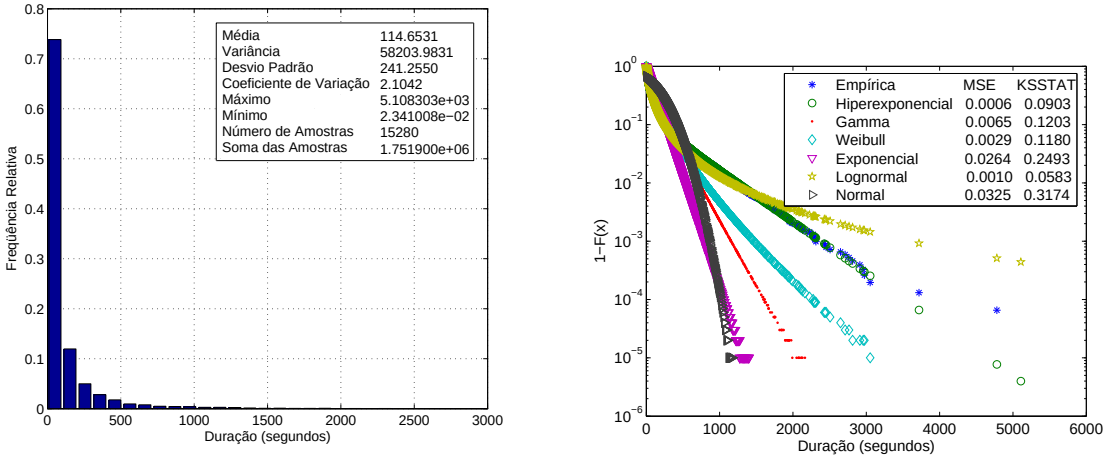
(a) (b)

Figura 3. (a) Frações de atividade e inatividade; (b) Frações de ocorrência de cada tipo de interação.

5.2. Tempo em *Play*

No gráfico com as CCDF's (Figura 4(b)) podemos perceber que a curva empírica está bem próxima da distribuição hiperexponencial, que parece capturar melhor o comportamento

dos dados. Os resultados de MSE e a estatística do teste de Kolmogorov-Smirnov confirmam este parecer. O QQplot das amostras empíricas versus as amostras geradas para uma hiperexponencial também indica que este deve ser um bom modelo para o tempo em *play*.



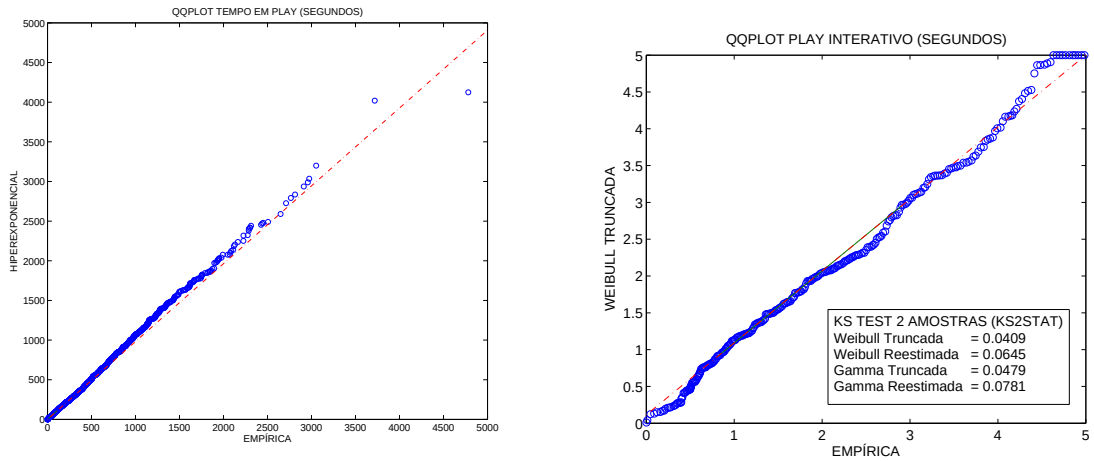
(a) (b)

Figura 4. (a) Histograma dos tempos em *play*; (b) Distribuição Cumulativa Complementar (eixo y em escala logarítmica).

Observamos que a grande maioria dos tempos em play são de curta duração, 84% das amostras são menores que 3 minutos. O valor médio de 114,65 segundos com coeficiente de variação igual a 2,1, indica alta interatividade dos usuários. A média do tempo em *play* encontrada nos trabalhos de [Tomimura et al. 2006] e [Padhye and Kurose 1998], 7,35 minutos e 8,3 minutos, respectivamente, estão bem acima da encontrada no nosso trabalho. O motivo para o valor baixo da média do tempo em *play* pode ser o conteúdo altamente interativo (com exercícios, animações das aulas, etc). Analisamos então separadamente sessões acima de 20, 30, 40 e 50 minutos. O objetivo foi observar se com sessões mais longas a média do tempo em *play* aumentaria. Mesmo com o uso desses filtros, a média do tempo em *play* não teve mudanças consideráveis. O maior valor médio encontrado foi 139,59 segundos com coeficiente de variação igual a 2,03, que ocorreu quando descartamos sessões menores que 40 minutos. Sendo assim, independente da duração das sessões, o tempo em *play* dos usuários do curso do CEDERJ é bem menor que o de usuários de outros sistemas analisados.

Para que a hiperexponencial se aproximasse da distribuição empírica de forma satisfatória, foram necessários 4 estágios. Os testes feitos com 2 estágios não revelaram um bom casamento entre a distribuição empírica e a hiperexponecial e uma melhora significativa não foi observada com aumento de 4 para 6 ou 8 estágios. Os parâmetros usados para a hiperexponencial são $\alpha = [0.2591, 0.1231, 0.2332, 0.3846]$ e $\lambda = [0.0515, 0.002, 0.0406, 0.0089]$, onde α_i é um elemento do vetor α, $0 \leq \alpha_i \leq 1$, $1 \leq i \leq 4$, e representa a probabilidade associada a cada estágio; e λ_i é um elemento do vetor λ, $1 \leq i \leq 4$, e representa a taxa associada a cada estágio.

As distribuições lognormal e gamma foram as que mais se aproximaram dos dados empíricos para o tempo em *play* no estudo de [Padhye and Kurose 1998]. Em



(a) (b)

Figura 5. (a) QQplot das amostras de *play*; (b) QQPlot das amostras de *play* interativo.

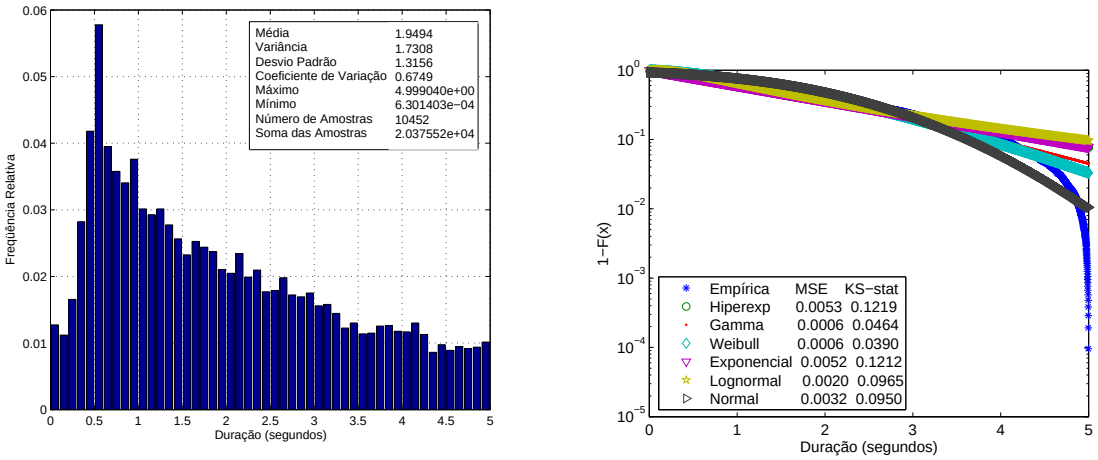
[Almeida et al. 2001], para arquivos de até cinco minutos, o tempo em *play* foi melhor aproximado por uma exponencial e por distribuições de cauda mais longa (pareto e weibull) para arquivos maiores. Já em [Costa et al. 2004] a pareto foi a que melhor representou o tempo em *play* para os arquivos de áudio e vídeo de curta duração, e para os vídeos de longa duração a distribuição weibull foi apontada como a melhor opção. As distribuições gamma, weibull e lognormal foram as que mais se destacaram em [Tomimura et al. 2006]. A distribuição escolhida pelos autores para representar o tempo em *play* no modelo proposto foi a lognormal truncada. (A distribuição foi truncada para que fosse considerado o tamanho dos vídeos.) A lognormal também foi a distribuição escolhida para o tempo que o usuário está assistindo continuamente um vídeo no sistemas de [Veloso et al. 2002] e [Branch et al. 1999].

5.3. *Play* Interativo

O tempos de play interativo, como descrito na seção 3, possuem valor máximo de cinco segundos, ou seja, a distribuição é truncada neste ponto (ver Figura 6(a)). O tempo médio calculado foi de 1.95 segundos, com coeficiente de variação igual a 0.67.

Os resultados dos testes apontam as distribuições gamma (*shape* γ = 1.7239 e *scale* α =1.1308) e weibull (*shape* γ =1.4532 e *scale* α =2.1446) como as mais adequadas para representar a métrica. Como esta é uma métrica truncada, ou seja, seu valor máximo é igual a 5 segundos, é preciso truncar as amostras geradas para as distribuições escolhidas. Foram geradas amostras aleatórias para as duas distribuições de duas formas: (i) todas as amostras geradas maiores que 5 segundos foram truncadas; e (ii) toda vez que uma amostra gerada é maior que 5 segundos, uma nova amostra é gerada. Este procedimento é repetido até que a amostra gerada seja inferior a 5s.

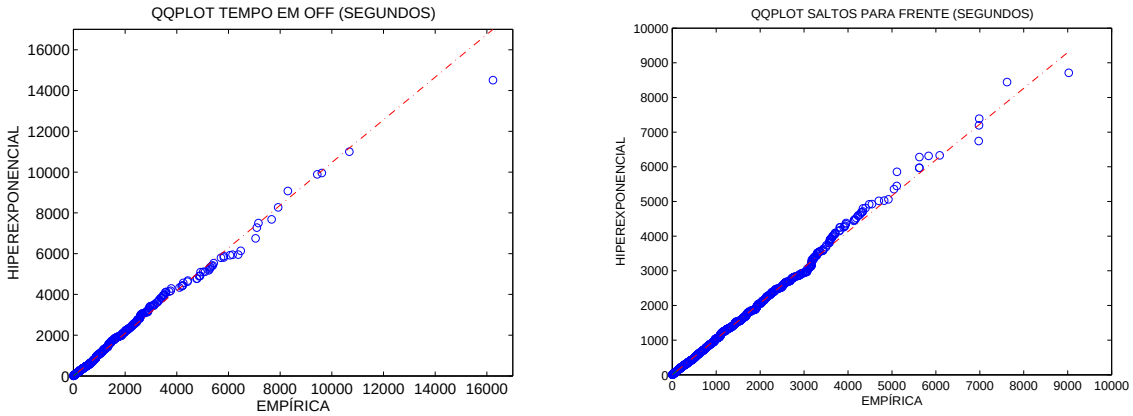
Para identificar qual a melhor das 4 opções descritas acima, usamos o teste de Komolgorov-Smirnov e o MSE. Os resultados dos testes apontaram a distribuição weibull truncada como a melhor opção para modelar o tempo em *play* interativo. O gráfico QQ-plot (Figura 5(b)) confirma a boa aproximação deste modelo com a curva da distribuição



(a) (b)

Figura 6. (a) Histograma dos tempos em *play* interativo; (b) Distribuição Cumulativa Complementar (eixo y em escala logarítmica).

empírica.



(a) (b)

Figura 7. (a) QQplot das amostras de tempo em OFF; (b) QQPlot das amostras saltos para frente.

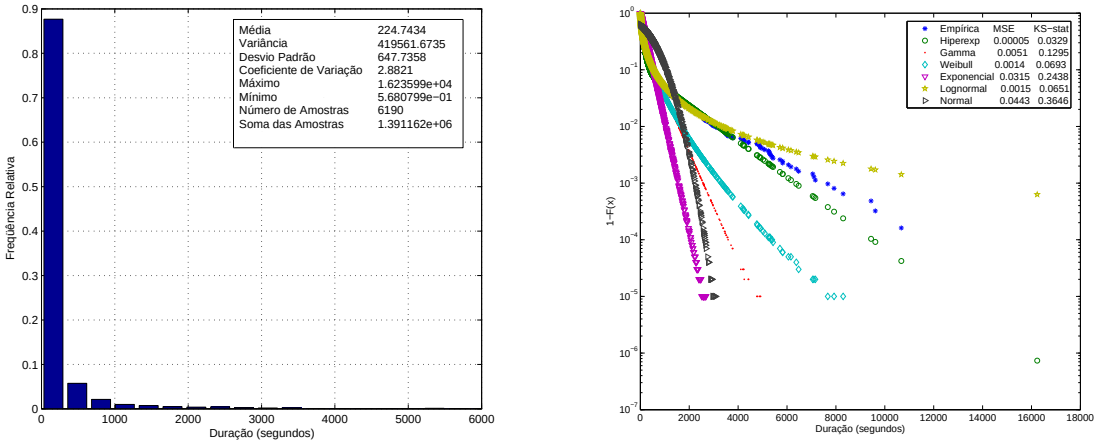
5.4. Tempo em OFF

Como aconteceu com a métrica tempo em *play*, a maior parte dos intervalos em *OFF* é de curta duração. A média em *OFF* foi de 224,74 segundos com coeficiente de variação igual a 2,9, sendo que 88% das amostras são menores que 5 minutos. Uma alta variabilidade nas amostras também ocorreu em [Tomimura et al. 2006]. O tempo médio em *OFF* encontrado foi de 1,6 minutos (96 segundos) com 2,5 de coeficiente de variação, sendo que 90% das amostras foram menores que 4 minutos. Vale ressaltar que no sistema estudado em [Tomimura et al. 2006] o tempo em *OFF* é composto pelos tempos gerados por outros tipos de interações, como *fast forward*, *fast rewind* e *index*, além dos tempos das interações de pausa e *stop*. Uma fração de 90% de amostras menores que 4 minutos se repete para os tempos em *OFF* do servidor eTeach em [Almeida et al. 2001]. Já o tempo

médio em *OFF* encontrado em [Padhye and Kurose 1998] foi de 171,91 segundos, com 392,24 de desvio padrão.

Contrariando nossas expectativas, os usuários do curso do CEDERJ passam, em média, mais tempo contínuo inativos do que assistindo efetivamente uma aula. Este fato pode ser explicado por uma característica única das aulas desenvolvidas para o curso: durante uma aula o professor pode solicitar que o aluno faça um exercício usando um outro programa qualquer ou desenvolver algum exercício proposto no vídeo. Neste momento o aluno vai pausar a aula. A parada pode representar também uma consulta ao livro texto.

As amostras, segundo os testes, são melhor representadas através da hiperexponencial. Quatro estágios foram necessários para que a hiperexponencial se ajustasse bem aos dados. Os resultados do ajuste podem ser vistos na Figura 8(b). Os vetores usados para parametrizar a hiperexponencial são: $\alpha = [0.1041, 0.0889, 0.491, 0.316]$ e $\lambda = [0.2781, 0.0007, 0.0124, 0.006]$.



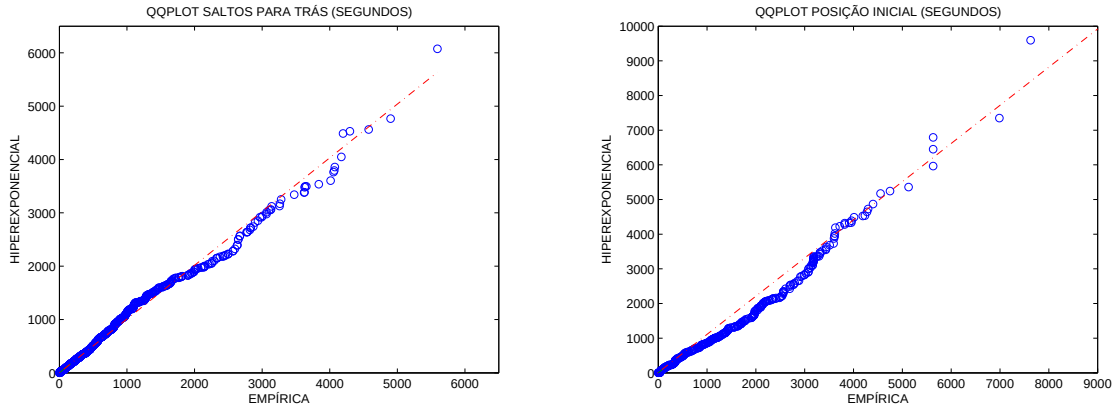
(a) (b)

Figura 8. (a) Histograma dos tempos em *OFF*; (b) Distribuição Cumulativa Complementar (eixo y em escala logarítmica).

Em [Padhye and Kurose 1998] a distribuição de melhor ajuste para o tempo em *OFF* foi a gamma, seguida pela lognormal. A distribuição weibull aparece como a melhor opção para todos os servidores e todas as durações de vídeos analisadas em [Costa et al. 2004]. Estas mesmas distribuições foram encontradas em [Tomimura et al. 2006], onde a distribuição escolhida para ser utilizada no modelo foi a lognormal. Em [Almeida et al. 2001] a exponencial foi a apontada para pequenos arquivos e a weibull, lognormal e pareto para arquivos maiores. A característica de cauda longa da distribuição dos tempos *OFF* também foi observada em [Veloso et al. 2002], onde é sugerido o uso de duas distribuições pareto com dois parâmetros diferentes.

5.5. Tamanho dos Saltos

Os saltos para trás foram 50% menos freqüentes que os saltos para frente, contrariando nossas expectativas de que o aluno retorne mais vezes no conteúdo das aulas para rever alguns tópicos. Os resultados encontrados indicam uma relativa localidade nos saltos para frente e para trás (maior ocorrência de saltos para posições próximas da atual).



(a) (b)

Figura 9. (a) QQplot das amostras de saltos para trás; (b) QQPlot das amostras de posição inicial acessada.

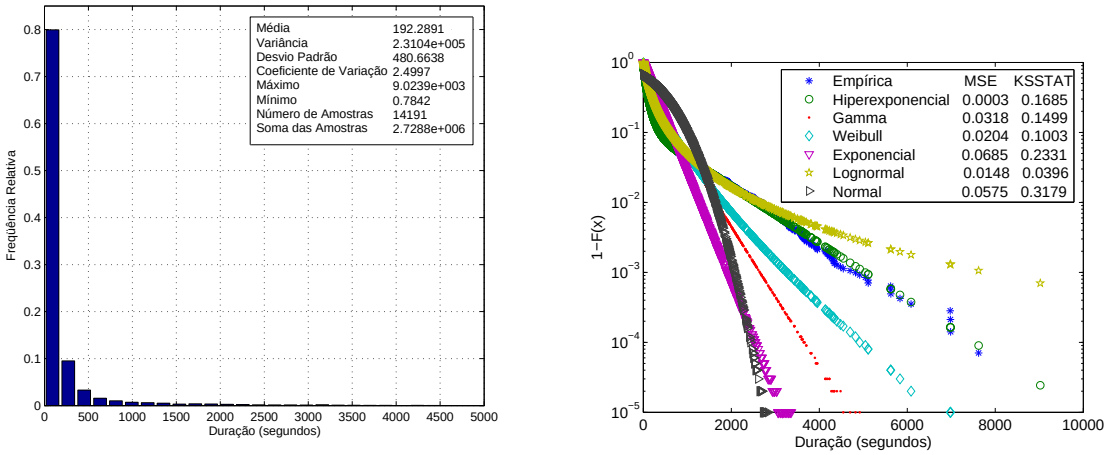
5.5.1. Saltos para frente

A maior parte das amostras corresponde a saltos para posições mais próximas a que o usuário se encontra, como pode ser observado na Figura 10(a). Obtivemos uma média igual a 192,29 segundos, coeficiente de variação igual a 2,5, e que 80% das amostras são menores que a média. O valor médio encontrado de saltos para frente em [Padhye and Kurose 1998] foi de 2137,62 segundos e cerca de 34,37% destes saltos foram menores que 3 minutos, indicando uma média bem superior a encontrada no nosso trabalho. Já o tamanho dos saltos para frente analisados em [Costa et al. 2004] possuem média bem inferior a encontrada no nosso trabalho. Para vídeos curtos, a média foi de 7 segundos e para vídeos acima de 5 minutos, a média foi de 40 segundos. Uma localidade temporal menor do que a encontrada no nosso trabalho e maior que a observada em [Costa et al. 2004], foi a obtida em [Tomimura et al. 2006], onde a média de saltos para frente foi de 82 segundos, com uma alta variabilidade (coeficiente de variação de 3,32), além de 87% das amostras serem menores que 180 segundos. Esta localidade dos saltos encontrada nos diversos sistemas estimula o uso de técnicas como *prefetching*, já que os saltos raramente são maiores que o valor médio encontrado.

A melhor distribuição para representar o tamanho dos saltos para frente foi a hiperexponencial com 6 estágios. O gráfico QQplot mostra quão bem a hiperexponencial se ajustou à curva das amostras empíricas (Figura 7(b)). Os vetores α e λ são, respectivamente: $\alpha = [0.1654, 0.203, 0.1093, 0.0915, 0.3252, 0.1055]$ e $\lambda = [0.0402, 0.0175, 0.0009, 0.01589, 0.0073, 0.0114]$. As distribuições gamma, lognormal e weibull foram apontadas em [Tomimura et al. 2006], como as melhores opções para caracterizar saltos para frente de até 500 segundos, referentes a 96% das amostras.

5.5.2. Saltos para trás

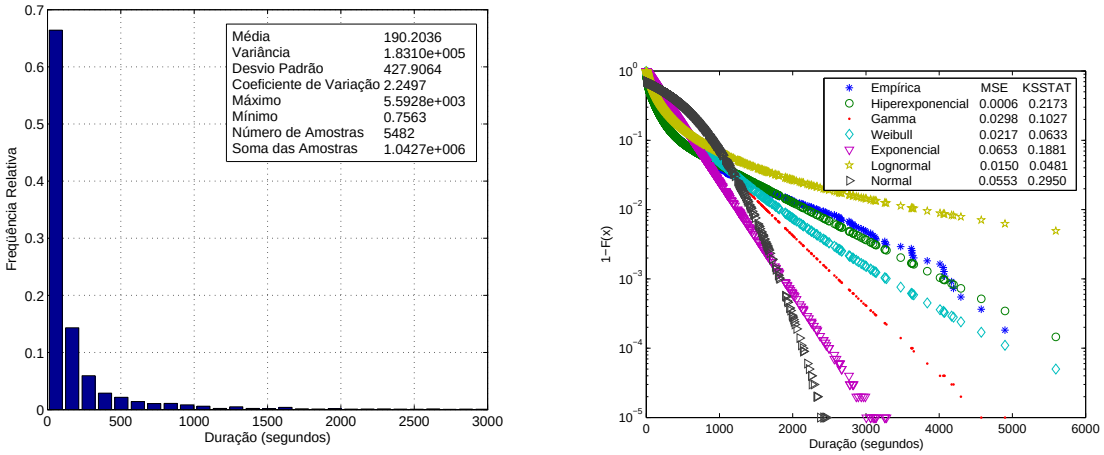
A média de tamanho dos saltos para trás foi de 192,2 segundos com coeficiente de variação igual a 2,25 e constatamos que 77% dos saltos para trás são menores que 3 minutos. O valor médio é novamente maior que os 143 segun-



(a) (b)

Figura 10. (a) Histograma dos saltos para frente; (b) Distribuição Cumulativa Complementar (eixo y em escala logarítmica).

dos encontrados em [Tomimura et al. 2006] e bem menor que os 3395 segundos de [Padhye and Kurose 1998]. Em [Costa et al. 2004], a média dos saltos para trás foi de 20 segundos, para vídeos curtos, e 40 segundos para vídeos maiores que 5 minutos. Como na análise dos saltos para frente, os resultados encontrados para os sistemas mostram que os saltos para trás também possuem alta localidade, possibilitando neste caso o uso de *buffers/proxies*, utilizando dados que já foram tocados e mantidos no cliente.



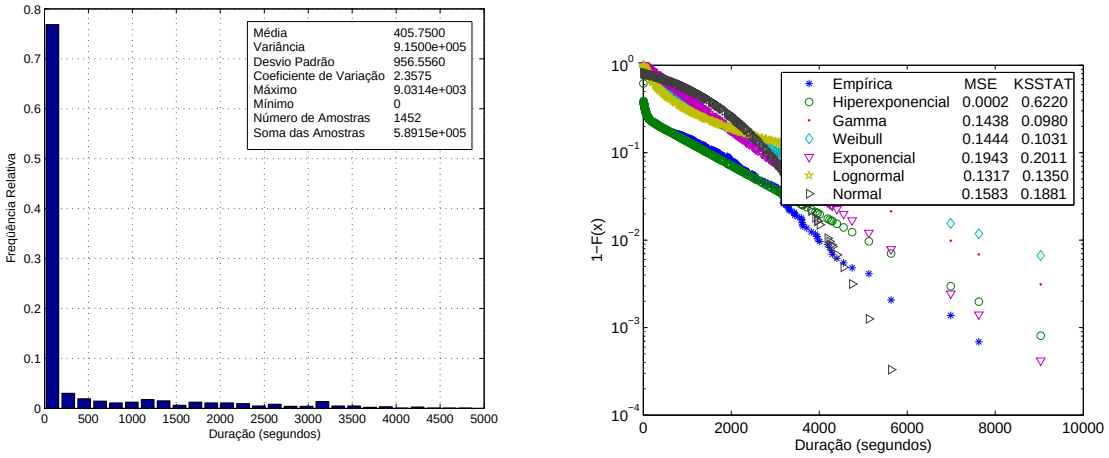
(a) (b)

Figura 11. (a) Histograma dos saltos para trás; (b) Distribuição Cumulativa Complementar (eixo y em escala logarítmica).

Novamente a distribuição hiperexponencial foi a que melhor se adequou a métrica. Os vetores de parametrização da hiperexponencial usada são: $\alpha = [0.3344, 0.1499, 0.3003, 0.2153]$ e $\lambda = [0.0653, 0.0012, 0.0094, 0.0068]$. Através do QQplot (Figura 9(a)) podemos ver o bom casamento das distribuições empírica e hiperexponencial.

5.6. Posição Inicial Acessada

A distribuição mais próxima da curva empírica, segundo o gráfico de CCDF's (Figura 12(b)), claramente é a hiperexponencial. Baseado nos valores de MSE e no gráfico QQplot, a melhor opção para esta métrica também é uma hiperexponencial com 4 estágios, com parâmetros $\alpha = [0.6006, 0.2534, 0.1257, 0.0202]$ e $\lambda = [0.9959, 0.0006, 0.0195, 0.0195]$.



(a) (b)

Figura 12. (a) Histograma da posição inicial acessada; (b) Distribuição Cumulativa Complementar (eixo y em escala logarítmica).

A média encontrada para a posição inicial foi de 405,75 segundos (com coeficiente de variação igual a 2,36), o que corresponde a iniciar a sessão aproximadamente no quinto slide. No entanto, em 60,8% das sessões o primeiro slide acessado foi o primeiro, conseqüentemente o bloco 0 (zero) foi o primeiro acessado nestes casos. Pelos dados acima podemos concluir que a maior parte dos alunos começa a assistir as aulas desde o início.

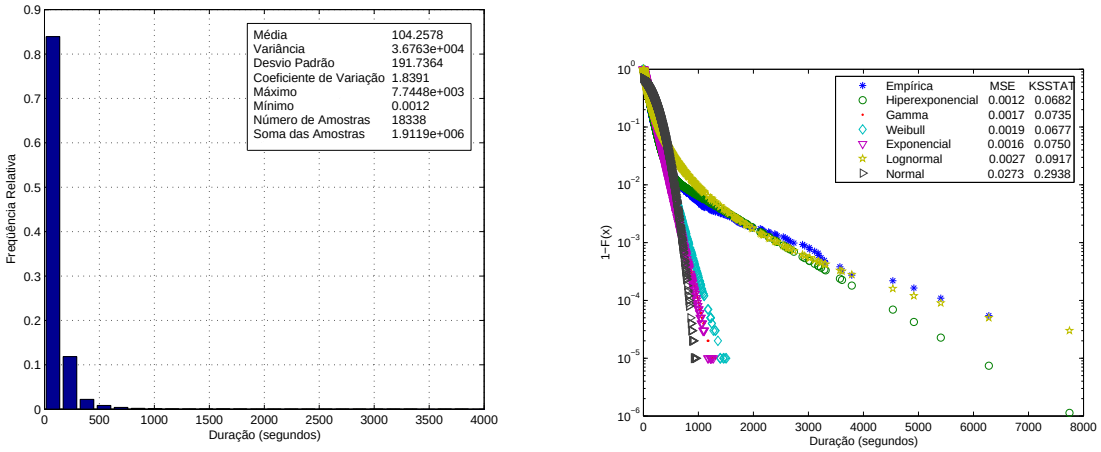
Em [Tomimura et al. 2006], 58% das sessão são iniciadas pelo primeiro slide da aula, sendo que os demais acessos se deram de forma uniforme entre as outras posições da aula. Portanto foi usada uma distribuição mista, onde os usuários iniciam uma aula no primeiro slide com 0,58 de probabilidade e nos demais slides de forma uniforme.

Para os servidores de áudio estudados em [Costa et al. 2004], quase todas as sessões iniciam pelo começo dos arquivos, nos servidores de vídeo uma quantidade significante de sessões inicia de outras posições do vídeo, quando se trata de arquivos maiores.

5.7. Permanência em um slide

A curva de distribuição complementar das amostras aparentemente tem a cauda acompanhada pela curva da distribuição lognormal conforme mostra a Figura 13(a). Entretanto até 2000 segundos, onde se concentra a grande maioria das amostras, a hiperexponencial parece se ajustar melhor. Os valores de MSE e KSSTAT apontam a hiperexponencial de dois estágios como a melhor opção de ajuste, o aumento no número de estágios não representou melhora no ajuste. O gráfico QQplot (Figura 14) mostra que as amostras geradas

para a hiperexponencial vêm de uma população semelhante a das amostras empíricas, indicando que esta distribuição é uma boa representação para esta métrica. Os parâmetros utilizados para a hiperexponencial foram $\alpha = [0.9768, 0.0232]$ e $\lambda = [0.0113, 0.0013]$.



(a) (b)

Figura 13. (a) Histograma do tempo de permanência em um slide; (b) Distribuição Cumulativa Complementar (eixo y em escala logarítmica).

O valor obtido para o tempo médio em um *slide* foi de 104,26 segundos e o coeficiente de variação foi de 1,8. Este valor médio representa 59,6% da duração do *slide*. Cerca de 43% dos alunos assistem um slide durante todo o tempo de sua duração, poucos acessos aos slides foram para tempos superiores à sua duração. Em uma análise semelhante em [Tomimura et al. 2006] a média encontrada foi de 138 segundos com coeficiente de variação 0,98. Em [Tomimura et al. 2006] as distribuições exponencial, weibull e gamma foram apontadas como as melhores aproximações para esta métrica.

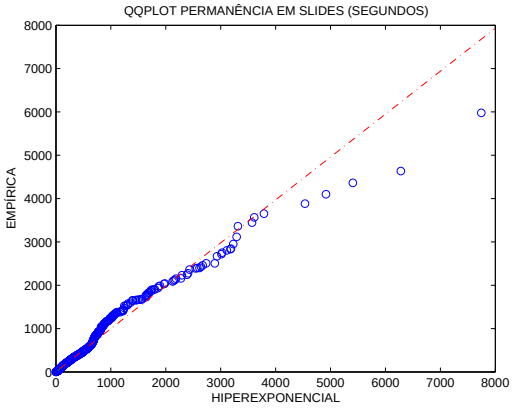


Figura 14. QQPlot das amostras de tempo de permanência em slide versus amostras geradas para a hiperexponencial.

A tabela 2 apresenta um resumo de algumas das métricas avaliadas nos trabalhos da literatura e no nosso trabalho.

6. Conclusões e Trabalhos Futuros

A análise dos logs apresentada revelou uma alta interatividade por parte dos usuários do sistema. O conteúdo bastante interativo das aulas do curso pode ter influenciado neste comportamento.

Como principais características do sistema podemos destacar:

- Uma fração grande das sessões são de curta duração (45% das sessões são menores que 5 minutos) e os vídeos de maior ocorrência no servidor e mais acessados são os de duração próxima a 50 minutos.
- As aulas introdutórias são mais acessadas que as outras do curso. Um decréscimo no número de acesso às aulas foi observado a medida que o conteúdo da disciplina é aprofundado.
- As interações mais freqüentes são as do tipo pausa e navegação pelo índice.
- Os saltos realizados pelos usuários apresentam localidade, o que pode beneficiar o uso de técnicas de melhoria de desempenho que aproveitam dados armazenados previamente.
- O fato de as aulas introdutórias serem as mais populares e os segmentos (blocos) iniciais das aulas serem os mais acessados, pode auxiliar na escolha de quais objetos e que parte deles devem ser armazenados previamente em um proxy intermediário para aumentar o desempenho do sistema.
- Grande parte das métricas pode ser modelada usando uma mesma distribuição (hiperexponencial) o que facilita a modelagem.

Como uma das principais contribuições deste trabalho, citamos o uso das distribuições obtidas para cada uma das métricas estudadas neste trabalho para parametrizar o modelo proposto em [Vielmond et al. 2007]. Este modelo está sendo usado no estudo de diversos mecanismos para o servidor, como protocolos de compartilhamento de banda e de gerenciamento de buffer. Outra aplicação deste modelo diz respeito a geração de carga para realização de testes visando avaliar o comportamento do servidor RIO distribuído em uma rede Gigabit/Ethernet (Projeto DIVERGE/GIGA).

Referências

Almeida, J. M., Krueger, J., Eager, D. L., and Vernon, M. K. (2001). Analysis of educational media server workloads. In *NOSSDAV '01: Proceedings of the 11th international*

Métrica	**ON (Play)**		***OFF***	
	Média	Distribuição	Média	Distribuição
RIO/CEDERJ	114,65	Hiperexponencial	224,74	Hiperexponencial
[Tomimura et al. 2006]	441	Lognormal	96	Lognormal
[Costa et al. 2004]		Pareto/ Weibull		Weibull
[Veloso et al. 2002]	166	Lognormal		Pareto
[Almeida et al. 2001]		Exponencial/Pareto/Weibull		Exponencial/Weibull Lognormal/Pareto
[Padhye and Kurose 1998]	509,50	Lognormal/Gamma	171,91	Gamma

Tabela 2. Comparação entre trabalhos (métricas em segundos).

workshop on Network and operating systems support for digital audio and video, pages 21–30, New York, NY, USA. ACM Press.

Branch, P., Egan, G., and Tonkin, B. (1999). Modeling interactive behaviour of a video based multimedia system. In *ICC '99: 1999 IEEE International Conference on Communications, 1999.*, pages 978–982.

Costa, C. P., Cunha, I. S., Borges, A., Ramos, C. V., Rocha, M. M., Almeida, J. M., and Ribeiro-Neto, B. (2004). Analyzing client interactivity in streaming media. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 534–543, New York, NY, USA. ACM Press.

dos Santos, J. R., Muntz, R. R., and Ribeiro-Neto, B. (2000). Comparing random data allocation and data striping in multimedia storage servers. In *ACM SIGMETRICS*, pages 44–55.

Netto, B. C. M., Azevedo, J. A., e Silva, E. A. S., and Leão, R. M. M. (2005). Servidor Multimídia RIO em Ensino a Distância. In *6th International Free Software Forum - Volume 1*, pages 91–95.

NIST. *DATAPLOT*. http://www.itl.nist.gov/div898/software/dataplot/.

NIST/SEMATECH (2006). *e-Handbook of Statistical Methods*. http://www.itl.nist.gov/div898/handbook/.

Olsson, M. The empht-programme.

O'Reilly and Associates. *Linguagem PERL*. http://www.perl.com/pub/q/documentation.

Padhye, J. and Kurose, J. (1998). An empirical study of client interactions with a continuous-media courseware server. In *Proc. 8th International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*.

Rocha, M., Maia, M., Cunha, I., Almeida, J., and Campos, S. (2005). Scalable Media Streaming to Interactive Users. In *MULTIMEDIA'05: Proceedings of the 13th annual ACM international conference on Multimedia*, Singapore.

The Mathworks. *MATLAB*. http://www.mathworks.com/.

Tomimura, D., Leão, R. M. M., de Souza e Silva, E., and Filho, F. S. (2006). Caracterização e modelagem do comportamento de usuários acessando um vídeo de ensino a distância. In *Anais XXVI Congresso da SBC - V Wperformance*, pages 34–53.

Trivedi, K. S. (2002). *Probability and statistics with reliability, queuing and computer science applications*. John Wiley and Sons Ltd., Chichester, UK.

Veloso, E., Almeida, V., Meira, W., Bestavros, A., and Jin, S. (2002). A hierarchical characterization of a live streaming media workload. In *IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pages 117–130, New York, NY, USA. ACM Press.

Vielmond, C., Leão, R. M. M., and de Souza e Silva, E. (2007). Um modelo HMM hierárquico para usuários interativos acessando um servidor multimídia. In *XXV Simpósio Brasileiro de Redes de Computadores (SBRC 2007)*.