Dimensionamento de Sistemas de Vídeo Interativo

Hana Karina Salles Rubinsztejn[1]

Instituto de Computação
Universidade Estadual de Campinas

Orientador

Prof. Dr. Nelson Luis Saldanha da Fonseca[2]

1 -CPqD Telecom & IT Solutions
Rod Campinas/Mogi-Mirim, Km 118,5
Caixa Postal 6070
13086-902 Campinas SP
Brasil
Tel: +55 19 37054404
hana@cpqd.com.br ou
hana.rubinsztejn@ic.unicamp.br

2 -Instituto de Computação
Universidade Estadual de Campinas
Caixa Postal 6176
13083-970 Campinas SP
Brasil
Tel: +55 19 37885878
Fax: +55 19 37885847
E-mail: nfonseca@ic.unicamp.br

Dimensionamento de Sistemas de Vídeo Interativo

Hana Karina Salles Rubinsztejn e Nelson L. S. da Fonseca
Universidade Estadual de Campinas
Instituto de Computação
Caixa Postal 6176
13083-970 Campinas SP
Brasil
E-mail: {nfonseca, hana.rubinsztejn}@ic.unicamp.br

Resumo

Para reduzir a grande demanda por banda passante dos serviços de vídeo sob demanda, técnicas baseadas em *multicast* são utilizadas, para o oferecimento de tais serviços em larga escala. Interatividade, uma característica desejável em serviços de vídeo, abrange a capacidade de efetuar operações de VCR. Sempre que um espectador solicita uma operação de VCR, seu fluxo de vídeo dessincroniza-se do fluxo de seu grupo de multicast. É necessário, portanto, o dimensionamento do número de canais necessários para a efetivação da interatividade em sistemas de vídeo-sob-demanda. Este artigo apresenta uma abordagem para determinar o número de canais de vídeo necessários em sistemas interativos. Além disso, o desempenho de sistemas interativos que possuem um conjunto de canais reservados para o suporte de operações de VCR, e que implementam *batching* e *piggybacking* é analisado.

Abstract

To reduce the hugh bandwidth demand of video-on-demand, techniques based on multicast have been considered for the deployment of such services in large scale. Interactiviness, a desirable feature of video services, encompasses the capability of performing VCR operations. Whenever a viewer issues a VCR operation, his/her video stream unsynchronizes with the stream of his/her multicast group. This paper introduces an approach to determine the number of video channels needed to support interactiviness. Moreover, the performance of interactive systems with reserved pool of channels for the support of VCR operations, as well systems with batching and piggybacking is analysed.

I - Introdução

Vídeo sob Demanda (*Video-on-Demand* - VoD) vem sendo utilizado em diversas áreas, tais como entretenimento, bibliotecas digitais e ensino a distância. Interatividade, uma característica desejável em serviços de vídeo sob demanda, é a capacidade de executar operações de VCR, tais como parada (PAUSE), retrocesso (REW) e avanço rápido (FF). Um sistema de vídeo sob demanda interativo (true video-on-demand) deve permitir requisições de operações de VCR a qualquer momento.

Para a manutenção de entrega contínua de um fluxo de vídeo, é necessário reservar recursos na rede e no servidor. Limitações de banda passante (*bandwidth*), de entrada/saída ou de capacidade de CPU, podem impor um limite no número de canais que um servidor é capaz de suportar. Em um sistema de VoD, se um canal de vídeo é atribuído a uma única requisição de exibição de filme, o número de usuários é limitado pelo número de canais do servidor. Entretanto, em redes com *multicast* pode-se prover um único canal a um grupo de usuários para assistirem simultaneamente a um filme. Mecanismos baseados em *multcast*, tais como *piggybacking* e *batching* [1], foram recentemente propostos, a fim de se reduzir a demanda por banda passante de serviços de vídeo em larga escala.

Piggybacking é baseado no fato de que espectadores não percebem uma variação na taxa de exibição de até 5% da taxa nominal [1]. Em um servidor de VoD com *piggybacking*, uma requisição para assistir a um filme é imediatamente aceita, se existir um canal disponível. Quando um novo canal é alocado e outra exibição do mesmo filme está em progresso, a taxa de exibição do filme em andamento é reduzida, enquanto que a taxa de exibição da requisição recem admitida é aumentada. Quando o fluxo de vídeo com maior taxa alcança o mais lento, os dois são mesclados e, conseqüêntemente, um dos canais de vídeo é liberado.

Em um servidor de VoD com *batching*, requisições para exibição de vídeo não são admitidas assim que chegam. Elas são atrasadas para que várias requisições de um mesmo filme dentro de um certo intervalo sejam coletadas. Um único canal de vídeo é, então, alocado para todo o grupo (*batch*) de requisições (usuários). Se, por um lado, *batching* aumenta a vazão, por outro, usuários podem não estar dispostos a esperar por longos períodos de tempo, e podem cancelar suas requisições (abandono). Sabe-se que um sistema com *batching* admite um maior número de usuários que um sistema com *piggybacking* [2]. Além disso, adotar ambos, *batching* e *piggybacking*, aumenta o número de usuários admitidos no sistema [3].

Quando um usuário efetua uma operação de VCR, sua exibição se dessincroniza com a exibição do seu grupo. Desta forma, um outro canal de vídeo é necessário para dar suporte ao fluxo dessincronizado. É desejável que um canal esteja imediatamente disponível para admitir a requisição de operação de VCR, de modo que seja assegurada a continuidade da exibição. Além disso, este usuário precisa de um canal dedicado até o final da exibição, ou até se ressincronizar com outro fluxo. Portanto, é de suma importância determinar o número de canais necessários para atender a demanda por banda passante das requisições de operações de VCR.

A principal contribuição deste artigo é uma nova técnica de dimensionamento do número de canais necessários, em um sistema de VoD interativo, para se prover a Qualidade de Serviço (*Quality-of-Service* - QoS) desejada. São introduzidas técnicas para sistemas de VoD com *batching*, e para sistemas de VoD com *batching* e *piggybacking*.

Este artigo está organizado da seguinte forma. Na seção II, o funcionamento de um sistema de VoD interativo é apresentado. Na seção III, o cálculo do número de canais necessários para suportar a demanda de operações de VCR é descrito. Na seção IV verifica-se a precisão do model

aproximado. Na seção V é introduzido um modelo para o sistema de VoD interativo. Resultados numéricos são discutidos na seção VI, e na seção VII conclusões são apresentadas.

II - Um Sistema de VoD Interativo.

Neste artigo, um sistema de VoD com *batching* e outro com *batching* e *piggybacking* são analisados. Requisições de exibição de vídeos não são imediatamente atendidas. Elas são agrupadas, e um único canal é alocado para um grupo de requisições, de acordo com uma política de *batching* específica. Além disso, em um sistema com *batching* e *piggybacking*, *n* fluxos de vídeo podem ser mesclados, liberando *n-1* canais, de acordo com o critério de piggybacking adotado.

Sempre que um usuário requisita uma operação de VCR, sua exibição se dessincroniza com a exibição de seu grupo multicast e, consequentemente, um canal dedicado deve ser alocado para ele. Dois esquemas são considerados neste artigo. No primeiro esquema, canais, chamados canais de contingência, são reservados para dar suporte a realização de operações de VCR. Sempre que um grupo de usuários é admitido no sistema, o número de canais no conjunto de contingência é calculado, de forma que possa acomodar futuras requisições de operações de VCR. Se não houver canais disponíveis no conjunto de contingência, um canal usado para *playback* pode ser alocado. O canal alocado é retido até o fim da exibição, ou até ocorrer uma mesclagem com outro fluxo.

No segundo esquema, não há reserva de canais para o suporte de fluxos de vídeo dessincronizados. Uma requisição de canal para realizar operações de VCR compete com requisições de admissão de novos grupos de usuários. Em ambos os esquemas, se o usuário que requisitou uma operação de VCR é o único usando o canal de vídeo, não há, obviamente, necessidade de se alocar um canal extra.

Nos dois esquemas, se não há canais disponíveis, a requisição é recusada, e o usuário permanece em seu grupo *multicast*. Em uma abordagem ortogonal poderia ser retardar a admissão da requisição. Esta abordagem não é considerada aqui, já que não há evidência clara que esta abordagem cause um impacto positivo na Qualidade de Serviço percebida pelo usuário. A banda passante alocada por um fluxo de vídeo para operações de VCR é a mesma alocada no modo de *playback*, isto é, quando REW e FF são executadas, a qualidade da imagem é reduzida.

III - Dimensionando o Número de Canais em um Sistema de VoD Interativo

Para assegurar o adequado funcionamento de um sistema de VoD interativo, o número de canais do conjunto de contingência deve ser dimensionado, de tal forma que apenas um pequeno número de requisições de operações de VCR sejam rejeitadas. Por isso, sempre que um usuário, ou grupo de usuários, é aceito no sistema, ou deixa o sistema, o número de canais do conjunto de contingência muda. Um usuário ou grupo de usuários deve ser aceito no sistema, se e somente se existir uma provisão de canais capaz de suportar o potencial número de operações de VCR a serem requisitadas.

Assume-se que a chegada de requisição de exibição de vídeo segue um processo de Poisson, e que os filmes são escolhidos de acordo com uma distribuição de Zip. Operações de VCR não são efetuadas por todos os usuários. Apenas aqueles que optem por estes serviços são permitidos emitir requisições de operações de VCR. O intervalo de tempo entre requisições de um mesmo usuário é exponencialmente distribuído, bem como a duração das operações de VCR. Estas suposições são fundamentadas em dados coletados em um sistema operacional [4].

A análise de um sistema com *batching* difere substancialmente da análise de um sistema com *batching* e *piggybacking*.

Um sistema com *Batching*

Em um sistema com *batching* sempre que um canal é alocado a um fluxo dessincronizado, ele é retido, até o fim da exibição, uma vez que não há como ressincronizar este fluxo de vídeo com outros fluxos. Portanto, dimensionar o número de canais do conjunto de contingência é uma tarefa simples. Toda vez que um grupo de *n* usuários é aceito no sistema, $n \times P_{u\_vcr}$ canais devem ser reservados para tratar os fluxos dessincronizados, sendo que P_{u_vcr} é a probabilidade de um usuário optar por serviços interativos.

Um Sistema com *Batching* e *Piggybacking*

Em um sistema com batching e piggybacking, após a execução de uma operação de VCR, o fluxo de vídeo pode se sincronizar com outro, permitindo a liberação do canal de vídeo. O tempo de retenção (*holding time*) de um canal de contingência inclui não só o tempo de execução da operação de VCR, mas também o tempo para se ressincronizar com outro fluxo.

O número de usuários no sistema altera-se nas admissões de grupos de usuários, e nos términos de exibições de vídeo. Portanto, entre a ocorrência de quaisquer dois destes eventos, o número de usuários permanece fixo. Assim, o sistema pode ser modelado por uma rede de filas fechada, ou seja, por um modelo de servidor central. No modelo de servidor central, o tempo em que usuários ficam no modo *playback* corresponde ao tempo de serviço de um servidor infinito (*Infinite Server*) [5]. Após visitar o servidor infinito, os usuários vão para uma fila com múltiplos servidores, isto é, uma fila M/M/c (um servidor dependente da carga - *load dependent server* na terminologia de redes de filas). O tempo de serviço desta fila corresponde ao tempo de retenção de um canal de contingência. Ele inclui o tempo para ressincronizar com outro fluxo, e o tempo médio de execução de uma operação de VCR, ponderados pela probabilidade da operação de VCR (PAUSE, REW ou FF). O número de servidores, *c*, é, na realidade, o tamanho do conjunto de contingência. O ponto chave é dimensionar *c*, de modo que nenhum usuário espere para ser servido. Sempre que um usuário achar todos os servidores ocupados, uma operação de VCR é rejeitada. Assim sendo, *c* tem que ser calculado de modo que leve a um baixo número de rejeições.

Para determinar o número de canais do conjunto de contingência, um algoritmo de redes de filas é executado com um certo valor de *c*, e o tamanho da fila do servidor dependente de carga é verificado. Este processo continua, até se encontrar um valor de *c* tal que o tamanho médio da fila no servidor dependente da carga seja muito pequeno. O modelo de servidor central é um modelo exato para o sistema de VoD interativo. No entanto, os algoritmos análise do valor médio (*Mean Value Analysis* - MVA) e de convolução [6], que são algoritmos exatos para redes de filas fechadas, apresentam instabilidades numéricas, quando o tamanho da fila nos servidores dependente da carga é muito pequeno. Alternativamente, o algoritmo de convolução normalizada [7] pode ser usado para superar tal instabilidade numérica. Entretanto, este algoritmo é também instável, quando o tempo de serviço da fila dependente de carga é uma parte significante do tempo de ciclo (*cicle time*), ou seja, o tempo para visitar o servidor infinito e a fila dependente de carga. Os algoritmos mencionados foram utilizados para resolver o problema proposto neste artigo, e a instabilidade numérica foi verificada, pois neste problema o tempo de serviço da fila dependente de carga compõe a maior parte do tempo de ciclo. Por exemplo, um usuário que permanece no estado de *playback* por 30 minutos e então requisita uma operação de PAUSE por 5 minutos terá um tempo de ressincronização de aproximadamente 50 minutos, ou seja, tempo de serviço de 55 minutos, sendo que o ciclo foi de 85 minutos.

Dada a instabilidade numérica dos algoritmos exatos para redes de filas fechadas, adotou-se um modelo aberto aproximado. O sistema de VoD é modelado por uma fila Erlang B. O servidor dependente de carga, nas redes de filas fechadas, corresponde a uma fila M/M/c sem fila de espera, no modelo aberto, isto é, requisições são perdidas se não houver servidor disponível. A taxa de chegada para o servidor dependente de carga, no modelo fechado, é calculada implicitamente. No entanto, a taxa média de chegada na fila M/M/c, no modelo aberto, precisa ser aproximada, como mostra a Figura 1. A aproximação é explicada a seguir.

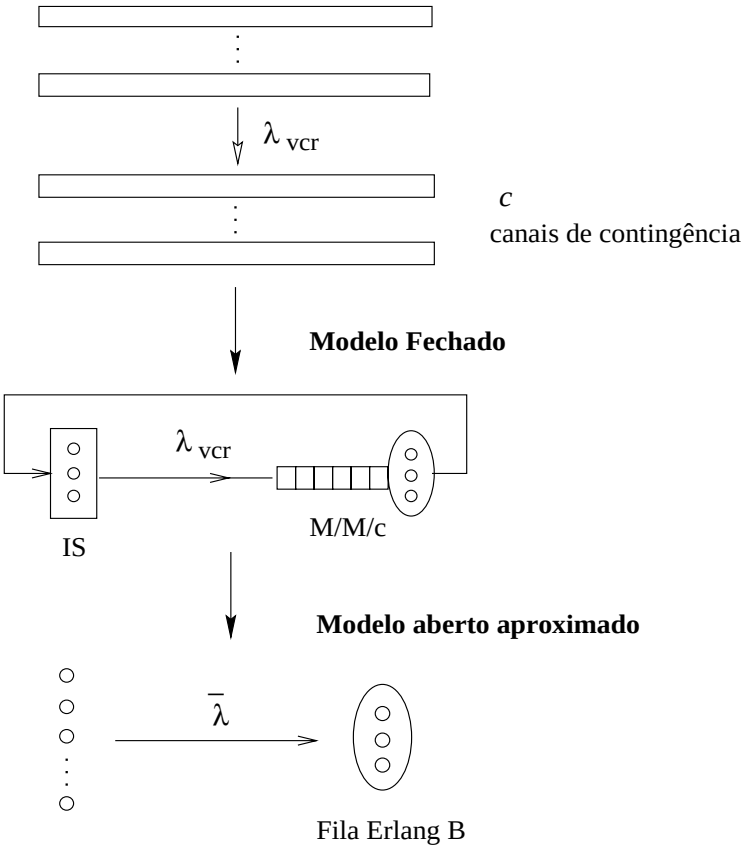

Figura 1: Aproximação do modelo de redes de filas fechadas para o modelo de filas aberto.

Um usuário pode estar em dois modos: *playback* ou VCR. No modo VCR, o usuário retém um canal de contingência e, no modo *playback,* ele é parte de um grupo *multcast*.

A taxa média de chegada de requisições VCR, isto é, a taxa média de chegada na fila M/M/c é a taxa de requisições de VCR por usuário multiplicada pelo número médio de usuários que efetuam operações de VCR.

$$\bar{\lambda} = N \ \lambda_{vcr} \ P_{playback}$$

onde N é o número de usuários que executam operações de VCR; λ_{vcr} é a taxa de requisições de VCR por usuário; $P_{playback}$ é a probabilidade de um usuário estar em modo de *playback*.

A probabilidade de estar em modo VCR, P_{vcr}, é a fração do tempo que um usuário retém um

canal de contingência durante toda a exibição. A duração da exibição é a duração original do filme adicionada do tempo gasto efetuando operações de VCR. Além disso, o tempo de retenção médio de um canal de contingência inclui o tempo médio de duração de uma operação de VCR e o tempo médio de ressincronização. P_{vcr} é então dado por:

$$P_{vcr} = (N_{vcr} \ D) / (T + N_{vcr} \ t_{vcr})$$

onde N_{vcr} é o número médio de operações de VCR efetuadas por um usuário; D é o tempo de retenção médio de um canal de contingência; t_{vcr} é a duração média de uma operação de VCR; T é a duração original do filme.

O tempo de retenção médio de um canal de contingência inclui o tempo médio de ressincronização com outro fluxo. Para simplificar o cálculo do valor de D, assume-se que o fluxo dessincronizado pode apenas mesclar-se com seu fluxo original. De fato, um fluxo pode mesclar-se com qualquer outro fluxo que esteja exibindo o mesmo filme. O valor calculado é, então, um limite superior para o valor de D, dado que o fluxo dessincronizado pode se mesclar com outro fluxo que esteja mais próximo do que seu fluxo original. O tempo de ressincronização depende do tipo de operação de VCR, de sua duração, e da posição do quadro na qual a operação de VCR foi requisitada. Além disso, a posição na qual a operação de VCR foi requisitada depende do número de operações realizadas durante a exibição do filme. D é, então, dado por:

$$D = \sum_{n=1}^{\infty} d(n) \ p(n)$$

onde $d(n)$ é o tempo de retenção médio de um canal de contingência dado que n operações são executadas durante a exibição do vídeo; $p(n)$ é a probabilidade de um usuário requisitar n operações de VCR durante a exibição do filme.

$p(n)$ e $d(n)$ são dados respectivamente por

$$p(n) = \frac{(\lambda_{vcr} \ T)^n}{n!} \ e^{(-\lambda_{vcr} \ T)}$$

$$d(n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{op} \sum_{s=i+1}^{L-i} d_{op}(s) \ \lambda_{vcr} \ \frac{\lambda_{vcr}^{(i-1)}}{(i-1)!} \ e^{-\lambda_{vcr}} \ P_{op}$$

onde λ_{vcr} é a taxa de requisições de operações de VCR de um usuário; L é o número de quadros do vídeo; $d_{op}(s)$ é o tempo de retenção do canal de contingência de uma operação de VCR op que ocorreu no s-ésimo quadro; P_{op} é a probabilidade do tipo da operação de VCR (PAUSE, FF, REW).

$\lambda_{vcr} \ \frac{\lambda_{vcr}^{(i-1)}}{(i-1)!} \ e^{-\lambda_{vcr}}$ - é a probabilidade de que a i-ésima operação seja requisitada no s-ésimo quadro.

Cada operação de VCR possui uma duração média própria. Desta forma, $d_{op}(s)$ deve ser computado como uma função do tipo da operação. Portanto

$$d_{op}(s) = \int_0^{Max_{op}(s)} G_{op}(s,t) \quad F_{op}(t) \quad dt$$

onde $Max_{op}(s)$ é a duração máxima da operação *op* que ocorreu no quadro *s*; $F_{op}(t)$ é a densidade de probabilidade da duração da operação *op*; $G_{op}(s,t)$ é a duração do tempo de retenção de um canal de contingência de uma operação de VCR que ocorreu no quadro *s* e durou *t* segundos.

Por exemplo,

$$G_{Rew}(s,t) = \begin{cases} t + (t + t\ R_{Rew}) \times A & se\ \ (t + t\ R_{Rew}) \times A \leq \dfrac{L - (s - t\ V_{Rew})}{V_{max}} \\ t + \dfrac{L - (s - t\ V_{Rew})}{V_{max}} & se\ \ (t + t\ R_{Rew}) \times A > \dfrac{L - (s - t\ V_{Rew})}{V_{max}} \end{cases}$$

onde $A = V_{playback} / (V_{max} - V_{min})$ é uma constante que, quando multiplicada pela duração da operação de VCR, produz o tempo necessário para se mesclar ao seu fluxo original; V_{max} e V_{min} - taxa máxima e mínima de exibição; $V_{playback}$ - taxa normal de exibição em *playback*; $V_{Rew} = R_{Rew} \cdot V_{playback}$; $V_{FF} = R_{FF} \cdot V_{playback}$ e *t* - duração da operação de VCR.

Por exemplo $d_{Rew}(s)$ é dado por:

$$d_{Rew}(s) = (1 + A \cdot e) \int_0^{Min(L'/(A\ e - B),\ s/V_{Rew})} t \quad \exp(\lambda_{Rew})\ dt \quad +$$

$$\int_{Min(L'/(A\ e - B),\ s/V_{Rew})}^{s/V_{Rew}} ((1 + B)\ t + L')\ \exp(\lambda_{Rew}) \quad dt$$

onde $L' = (L - s)/V_{Max}$, $B = V_{Rew}/V_{max}$, $e = 1 + R_{Rew}$.

IV -Precisão do Modelo Aproximado

Para avaliar a precisão do modelo aproximado, o número estimado de canais demandados foi confrontado com resultados derivados via simulação. Ao invés de simular todo o sistema, apenas a chegada de requisições de operações de VCR ao conjunto de contingência foi simulada, visto que isto é o que realmente influencia o valor estimado. Para variar $\bar{\lambda}$, que representa a taxa média de chegada de requisições de operações de VCR, diferentes valores de *N* e λ_{vcr} foram escolhidos. Desde que $\lambda_{vcr} = N_{vcr} / T$, diversos valores de λ_{vcr} foram obtidos através da variação de N_{vcr}. Variou-se *N* no intervalo [50, 2000] e N_{vcr} em [1,5]. Diferentes valores

de $P_{playback}$ foram obtidos alterando a duração média das operações de VCR, isto é, a média de $F_{op}(t)$.

O número de canais necessários estimado pelo modelo aproximado é um limite superior do número médio de canais demandados, e é um limite inferior do número máximo de canais demandados.

Assume-se que um fluxo de vídeo que retém um canal de contingência pode apenas se ressincronizar ao fluxo associado ao seu grupo *multicast* original. Esta suposição superestima o tempo médio de retenção dos canais de contingência e, conseqüentemente, leva a uma estimativa conservadora do número requerido de canais de contingência.

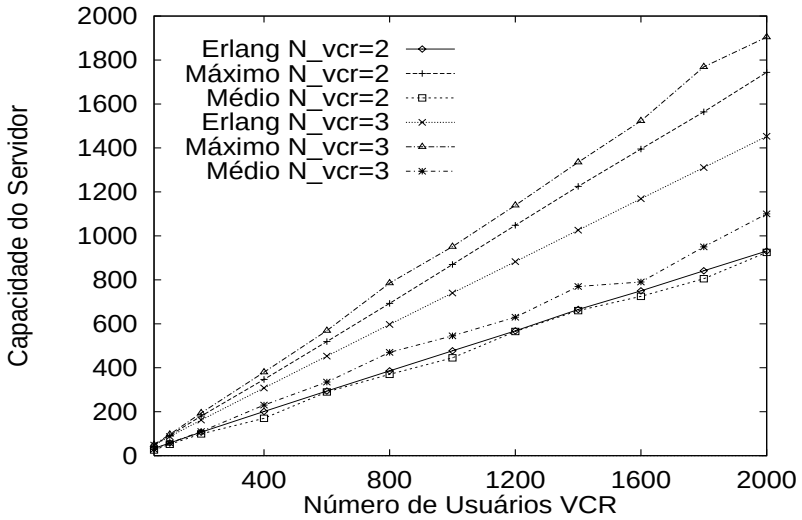

Figura 2: Comparação entre o número de canais de contingência estimado e o número médio e máximo de canais obtidos via simulação.

A Figura 2 ilustra o número de canais estimado, o número médio e o número máximo de canais obtidos via simulação, em função de *N*, para diferentes valores de N_{vcr} . N_{vcr} é o parâmetro de maior influência para que o valor estimado seja mais próximo ao valor médio ou ao valor máximo. Quanto maior o valor de N_{vcr} , mais próximo do valor máximo está o valor estimado. Esta tendência pode ser entendida pelo fato de que quanto maior o número de visitas ao conjunto de contingência, maior é a possibilidade da requisição de operação de VCR ser emitida em uma posição próxima ao final do vídeo, o que não permite futuras mesclagens com o fluxo *multicast* original. Conseqüentemente, maior é o tempo de retenção médio de um canal de contingência.

A fórmula de Erlang é calculada sempre que um grupo de usuários é admitido no sistema, ou seja, ela tem que ser calculada em tempo real. A tabela 1 apresenta o tempo médio para se calcular a fórmula de Erlang, para diferentes valores de *N* e N_{vcr} . Estes tempos foram obtidos em uma máquina Sparc com sistema operacional SunOS e 512 Mb de memória.

Número de Usuários VCR	Tempo Médio de Execução (μs)			
	Número de Operações de VCR por filme			
	1	2	3	4
50	5	10	10	10
100	10	15	20	30
400	20	50	70	90
1000	45	110	170	220
1600	70	180	260	350
2000	100	220	330	420

Tabela 1: Tempo médio do cálculo da fórmula de Erlang, em microsegundos.

V - Um Modelo de Sistema VoD

Para se comparar diferentes sistemas de VoD, um simulador de um sistema de VoD interativo foi desenvolvido. Os parâmetros de desempenho nos quais se tem interesse são *i*) o número de usuários admitidos no sistema, *ii*) a probabilidade de abandono, isto é, a razão entre o número de usuários que desistem de assistir ao filme e o número total de requisições que chegaram no sistema, e *iii*) a porcentagem de operações VCR rejeitadas. Provedores de serviço de VoD almejam aumentar o número de usuários admitidos no sistema, isto é, o número de requisições de exibição de filme aceitas. A taxa de abandono traduz o prejuízo causado pela não admissão de requisições de exibição de vídeo. A porcentagem de operações de VCR rejeitadas representa a porcentagem de requisições de operações de VCR que são recusadas, devido a falta de canal para o seu suporte. Ela representa o grau de quebra de contrato entre usuários e o provedor de serviço, e é uma medida da qualidade do serviço fornecido.

Chegadas de requisições seguem um processo de Poisson com média λ. Ao chegar uma requisição, o vídeo associado a ela é escolhido de acordo com a distribuição de Zip, e a requisição entra na fila específica. A distribuição de Zip modela de forma precisa a preferência de usuários em locadoras de vídeo. De tempos em tempos, uma decisão é tomada sobre a qual filme um canal deve ser alocado, de acordo com a política de *batching* adotada. Uma vez que nem todos os usuários emitem requisições de operações de VCR, quando um grupo de usuários é aceito no sistema, isto é, um canal é alocado a eles, o número de usuários naquele grupo que efetuam operações de VCR é determinado randomicamente. A probabilidade de um usuário executar operações de VCR é dado por p_{u_vcr}. O número médio de operações de VCR efetuadas por um usuário durante a exibição de um vídeo é dado por N_{vcr} operações. O intervalo de tempo entre requisições de operações de VCR é obtido de uma distribuição exponencial com média dada pela duração do vídeo dividida pelo número médio de operações efetuadas por um usuário, N_{vcr}. O tipo da operação de VCR é determinado por p_{op} e sua duração é exponencialmente distribuída com média λ_{op}, onde *op* pode ser REW, FF ou PAUSE.

A política de *batching* usada no experimento de simulação é a *Look-ahead-maximize-batch*

(LAMB) [2] e a política de *piggybacking* é a par-ímpar [1]. LAMB foi escolhida porque admite o maior número de usuários quando comparada a outras políticas de *batching*. Quanto maior o número de usuários admitidos no sistema, para um valor fixo de p_{u_vcr}, maior será o número de usuários que requisitam operações de VCR. Conseqüentemente, uma avaliação mais precisa da abordagem proposta é obtida. Sob LAMB, pontos de escalonamento são definidos pelo tempo de abandono de usuários. LAMB maximiza o número de usuários admitidos em uma janela de tempo definida pelo tempo de escalonamento e o tempo de abandono mais longe de um usuário no sistema. LAMB leva em consideração todas as liberações de canais dentro desta janela de tempo. Quando a decisão de aceitar um novo grupo de usuários é tomada, o número de canais extras necessários para suportar a demanda por operações de VCR deste novo grupo de usuários é calculado, usando a aproximação desenvolvida na seção III. Então, a disponibilidade de tal número de canais é verificada. Se o número de canais disponíveis não é suficiente para suportar o novo grupo, este não é admitido no sistema neste momento. Caso contrário, o número de canais do conjunto de contingência é ajustado para o valor corrente, adicionado dos canais extras exigidos pelo novo grupo.

Adotou-se a política de *piggybacking* par-ímpar, que mescla fluxos em pares. Algumas políticas, como Snapshot [8] e S2 [9], produzem um menor número de quadros exibidos do que a política adotada, entretanto assumem um processo de chegada de Poisson, suposição que não é válida em um sistema sob batching. As políticas par-ímpar e mesclagem simples produzem os mesmos resultados quando usadas em conjunto com LAMB [3].

VI - Resultados

Para avaliar o desempenho do sistema de VoD, utilizou-se simulação de eventos discretos. O método de replicação independente foi utilizado para derivar intervalos de confiança com 99% de nível de confiança. O número de replicações para cada ponto das curvas exibidas nesta seção foi tal que a largura dos intervalos de confiança é de no máximo 8% do valor médio. Intervalos de confiança não são exibidos nos gráficos para uma melhor interpretação visual dos resultados. Sistemas de VoD com batching e sistemas com batching e piggybacking são estudados. A eficiência da adoção de canais de contingência é também investigada [10,11].

Foram realizadas simulações para variadas configurações, graus de interatividade e sob diferentes cargas. Os resultados são apresentados em função da capacidade do sistema, isto é, número de canais disponíveis para transmitir fluxos de vídeo. O impacto do tamanho do servidor, ou seja, o número de vídeos armazenados, no desempenho é também avaliado. Para uma carga fixa, variações na taxa de requisições de operações de VCR foram obtidas através alterações no grau de interatividade, isto é, na fração de usuários que fazem operações de VCR, p_{u_vcr}. O valor desejado para a probabilidade de abandono, bem como para a rejeição de operações de VCR, é 1%.

Inicialmente são apresentados os resultados para servidores pequenos, com 100 filmes armazenados com duas horas de duração cada. Também são mostrados resultados para baixas cargas de 10 requisições por minuto, e para cargas elevadas de 60 requisições por minuto. A probabilidade de uma requisição de PAUSE é 0.5, e sua duração média é 5 minutos. As probabilidades de FF e REW são ambas de 0.25 e sua duração média é 30 segundos. O número médio de operações de VCR por usuário é 2. A sensibilidade do desempenho do sistema ao comportamento do usuário é descrita ao final desta seção.

Sempre que um grupo de usuários é admitido no sistema, o número de canais necessários para operações de VCR é calculado utilizando o modelo aproximado introduzido na seção III.

Sistema de VoD com *Batching*

Figuras 3, 4 e 5 mostram, respectivamente, o número de usuários admitidos no sistema, a probabilidade de abandono e a porcentagem de operações de VCR rejeitadas em função, da capacidade do servidor, para diferentes graus de interatividade, e para um servidor pequeno. Em todas as figuras aqui exibidas, CC e U denotam "com canais de contingência" e grau de interatividade, respectivamente.

O número de usuários admitidos no sistema é sempre maior para um servidor sem o conjunto de canais de contingência do que para servidores com este conjunto. Para um grau de interatividade de 10%, não há diferença entre os sistemas com e sem conjunto de contingência. Entretanto, conforme o grau de interatividade aumenta, a diferença entre o número de usuários admitidos em ambos os sistemas aumenta. Para um grau de 80%, e para um sistema com capacidade de 1000 canais, o número de usuários admitidos no sistema sem conjunto de contingência é maior que o dobro do valor do número de usuários admitidos em um sistema com conjunto contingência. Esta diferença diminui conforme a capacidade do sistema aumenta, pois o número de usuários admitidos no sistema converge para seu valor máximo, para uma carga fixa. Em um sistema com conjunto de contingência, canais permanecem ociosos para suportar operações de VCR. Tais canais são utilizados para admitir mais usuários no sistema sem canais de contingência.

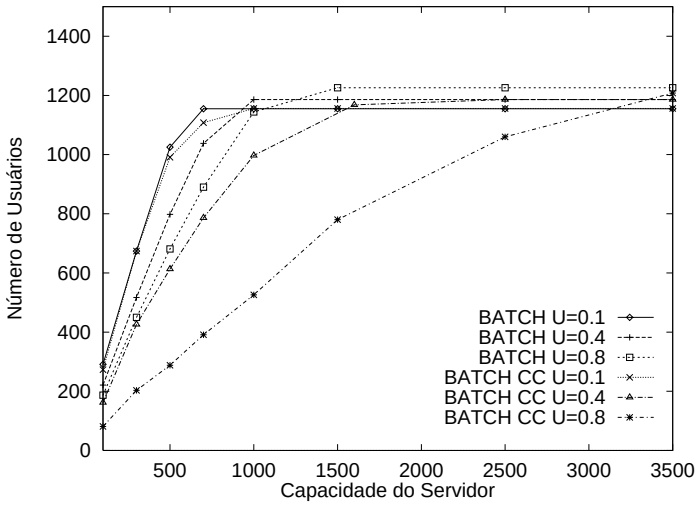

Figura 3: Número de Usuários Admitidos X Capacidade do Servidor, taxa de 10 req/min.

Note que o número de usuários admitidos no sistema não cresce significativamente apesar do aumento da capacidade do sistema. Este efeito de limite superior aparece em todos os tipos de sistemas investigado neste estudo, e pode ser atribuído ao comportamento regulador das políticas de *batching*. Sob *batching*, os usuários são admitidos de maneira discreta, e não de forma contínua. Usuários são admitidos de tempos em tempos, e o tempo entre admissões é determinado pela política de *batching* adotada. Então, não é vantajoso aumentar a disponibilidade de canais além de um certo valor já que a demanda por novos canais é regulada pelo intervalo entre admissões da política de *batching* . Por exemplo, em um sistema com canais de contingência e com grau de interatividade de 40\%, o número de usuários admitidos permanece constante após 1600 canais disponíveis.

A probabilidade de abandono para sistemas sem conjunto de contingência é sempre menor que para sistemas com conjunto de contingência (Figura 4). A diferença entre o número de canais necessários para suportar uma probabilidade de abandono de 1%, por um sistema com conjunto de contingência e por um sistema sem este conjunto, aumenta conforme o grau de interatividade aumenta. Para um baixo grau de interatividade (10%), esta diferença é de 40%, enquanto que, para um grau de interatividade de 80%, o número de canais demandados por um sistema com conjunto de contingência é o dobro do número de canais demandados por um sistema sem conjunto de contingência.

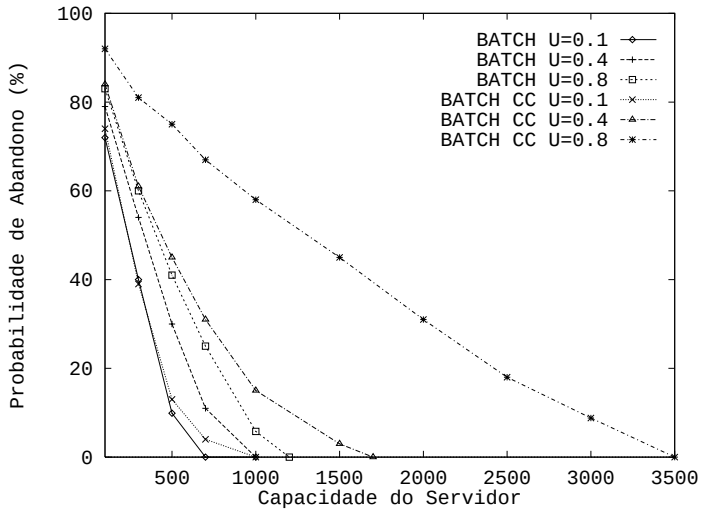

Figura 4: Probabilidade de Abandono X Capacidade do Servidor, taxa de 10 req/min.

Para uma capacidade de sistema acima de 300 canais, o número de operações de VCR rejeitadas em um sistema com canais de contingência está abaixo de 1%. Por outro lado, para um grau de interatividade de 10%, em um sistema sem canais de contingência, tal porcentagem é atingida com uma capacidade de sistema de 700 canais. Para um grau de interatividade de 80%, esta porcentagem é obtida para uma capacidade de sistema de 1.300 canais. Apesar disso, sistemas com canais de contingência não são preferíveis sobre sistemas sem canais de contingência . Note que, para graus de interatividade baixos (10%), em um sistema sem conjunto de contingência, 700 canais são necessários para prover uma porcentagem de requisições de VCR rejeitadas de 1%, enquanto que em um sistema com conjunto de contingência, 1.000 canais são necessários para obter uma probabilidade de abandono de 1%. Em outras palavras, os valores de desempenho desejados são obtidos com uma menor capacidade em sistemas sem conjunto de contingência do que em sistemas com conjunto de contingência. Para elevados graus de interatividade (80%), em um sistema sem canais de contingência, a porcentagem de requisições de VCR rejeitadas, bem como a probabilidade de abandono, estão abaixo dos valores alvo para uma capacidade de sistema de 1.400 canais, enquanto que em um sistema com conjunto de contingência, a probabilidade de abandono está abaixo do valor alvo para uma capacidade de sistema de 3500 canais. Portanto, para elevados graus de interatividade, sistemas sem canais de contingência possuem um desempenho melhor que sistemas com canais de contingência.

Sob altas cargas (60 requisições por minuto), a diferença entre o número de usuários admitidos em um sistema sem conjunto de contingência e em um sistema com conjunto de contingência

é superior a diferença sob baixas cargas, como ilustrado na Figura 6. Tal diferença aumenta conforme o grau de interatividade aumenta. Por exemplo, para 1000 canais, e grau de interatividade de 10%, a diferença é 300 usuários, enquanto que, para um grau de interatividade de 80%, ela é 1.100 usuários.

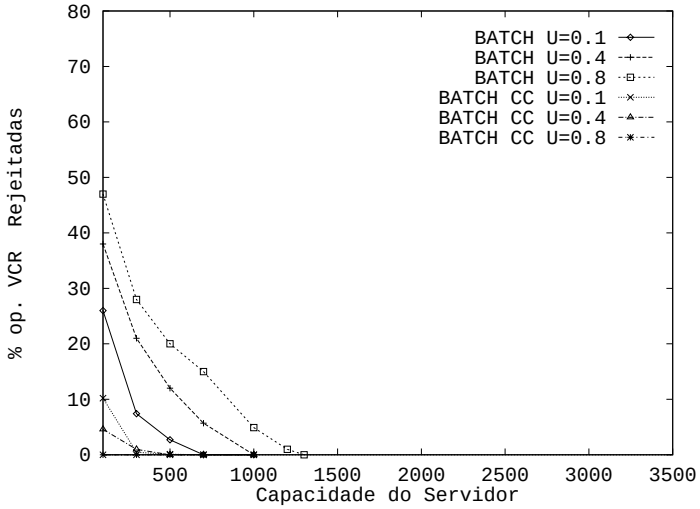

Figura 5: Porcentagem de op. VCR Rejeitadas X Capacidade do Servidor, taxa de 10 req/min.

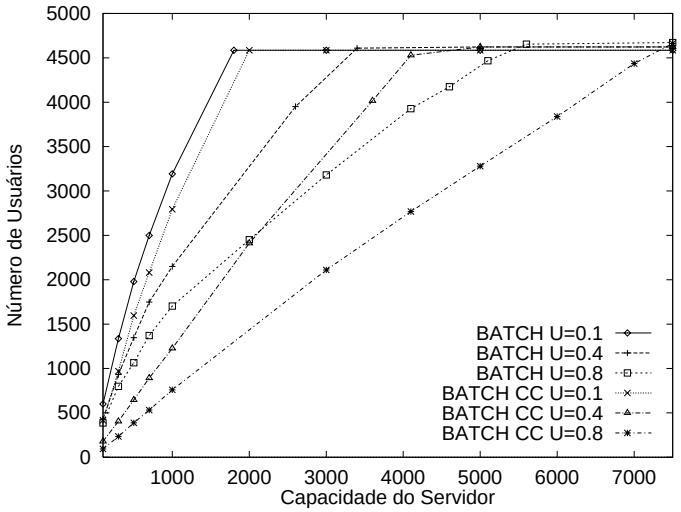

Figura 6: Número de Usuários Admitidos X Capacidade do Servidor, taxa de 60 req/min.

Para atingir a meta de probabilidade de abandono, um maior número de canais é necessário que sob uma carga baixa, como pode ser visto na Figura 7. A probabilidade de abandono para sistemas com conjunto de contingência é maior que para sistemas sem conjunto de contingência, sem levar em consideração o grau de interatividade. No entanto, a diferença entre as probabilidades de abandono é menor que sob baixas cargas. Sob altas cargas, o valor alvo para a porcentagem de operações de VCR rejeitadas é atingido com uma capacidade de sistema que é o dobro da capacidade necessária sob baixas cargas, para um sistema sem canais de contingência (Figura 8). Por exemplo, para 80%

de interatividade, o valor alvo é atingido com 1250 canais sob baixas cargas, enquanto que sob altas cargas, são necessários 5600 canais.

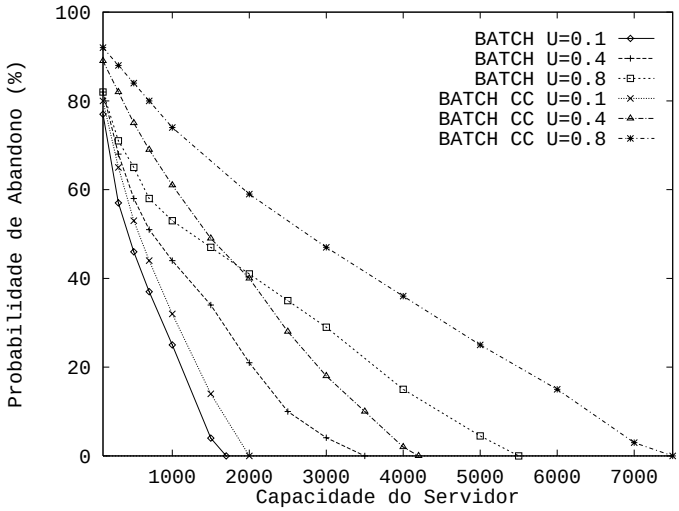

Figura 7: Probabilidade de Abandono X Capacidade do Servidor, taxa de 60 req/min.

Embora um maior número de canais sejam demandados para se atingir a porcentagem alvo de requisições de operações de VCR rejeitadas para um sistema sem canais de contingência, o número de canais exigidos por um sistema com canais de contingência para atingir a probabilidade alvo de abandono (Figura 7) é maior que o número de canais para atingir os valores alvos de desempenho por um sistema sem canais de contingência. Por exemplo, para um grau de 40% de interatividade, um sistema sem canais de contingência necessita de 3500 canais para atingir suas metas, enquanto que um sistema com canais de contingência precisa de 4100 canais.

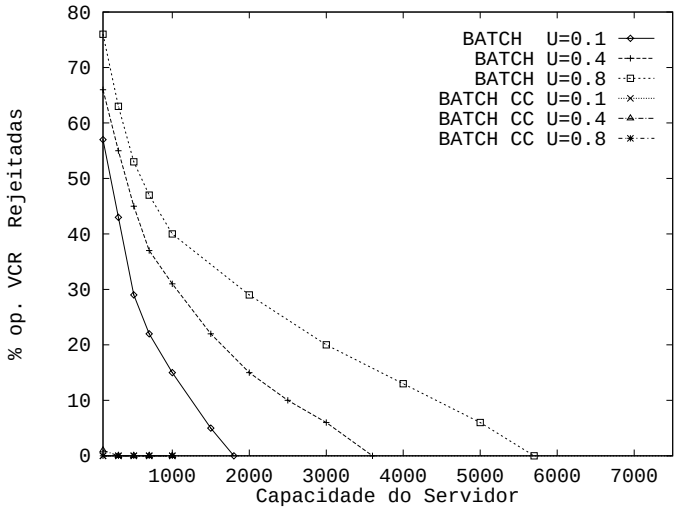

Figura 8: Porcentagem de op. VCR Rejeitadas X Capacidade do Servidor, taxa de 60 req/min.

Para um servidor grande (500 filmes), o número de canais necessários para obter a probabilidade de abandono desejada é tipicamente maior que para um servidor pequeno. Isto ocorre porque

probabilisticamente há um maior número de canais alocados a um único usuário em um sistema grande do que há em um servidor pequeno.

Sistemas de VoD com *Batching* e *Piggybacking*

Figuras 9, 10 e 11 apresentam, respectivamente, o número de usuários admitidos no sistema, a probabilidade de abandono e a porcentagem de operações de VCR rejeitadas, para um servidor pequeno, sob cargas baixas. Figuras 12, 13, 14 mostram os mesmos gráficos para um servidor pequeno, sob cargas altas.

Sob cargas baixas, o número de usuários admitidos no sistema com *batching* e *piggybacking* segue a mesma tendência encontrada em um servidor com *batching* apenas. A diferença entre o número de usuários admitidos em um sistema com conjunto de canais de contingência, e o número de usuários admitidos em um sistema sem conjunto de contingência aumenta com o grau de interatividade. No entanto, em um sistema com *batching* e *piggybacking*, esta diferença é menor no que em um sistema com *batching*, para graus de interatividade altos, dado que o número de usuários admitidos converge mais rapidamente para o seu valor máximo. Por exemplo, para 80\% de interatividade e 1000 canais esta diferença é de 650 usuários em sistemas com *batching*, enquanto que para sistemas com *batching* e *piggybacking* ela é de 200 usuários. Para um grau de interatividade baixo esta diferença não é significativa.

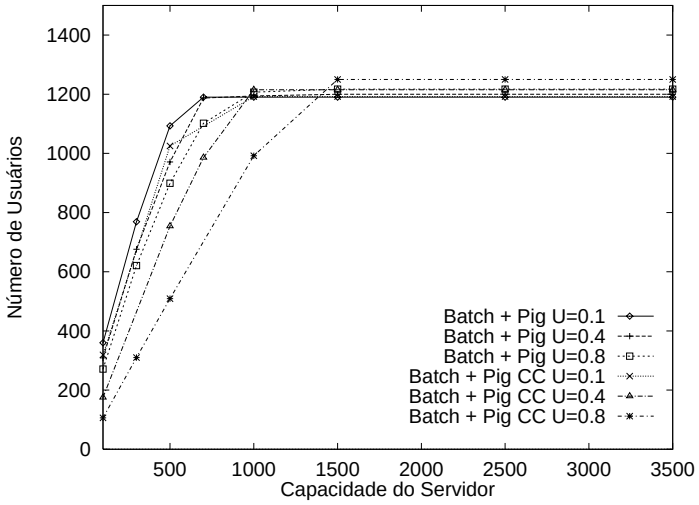

Figura 9: Número de Usuários Admitidos X Capacidade do Servidor, taxa de 10 req/min.

Para um grau de interatividade fixo, sistemas sem conjunto de canais de contingência fornecem uma probabilidade de abandono menor que sistemas com conjunto de canais de contingência (Figura 10). Entretanto, em um sistema com *batching* e *piggybacking*, valores específicos de probabilidade de abandono são obtidos com um número menor de canais do que em um sistema com apenas *batching* (Figura 4). Para sistemas com canais de contingência e para graus de interatividade médio e alto, necessita-se do dobro do número de canais em um sistema com batching, para se obter um valor da probabilidade de abandono de 1%, do que se necessita em um sistema com *batching* e *piggybacking*. Por exemplo, em um sistema com canais de contingência para 40% de interatividade, a probabilidade de abandono de 1% é atingida com 850 e 1700 canais em sistemas com *batching* e *piggybacking* e em sistemas com apenas *batching*, respectivamente.

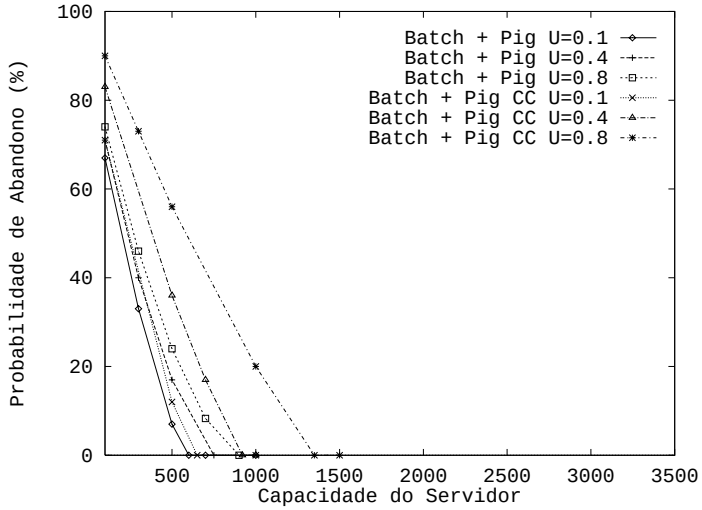

Figura 10: Probabilidade de Abandono X Capacidade do Servidor, taxa de 10 req/min.

O mesmo padrão da porcentagem de operações de VCR rejeitadas observado em um sistema com *batching* apenas (Figura 5) é observado em um sistema com *batching* e *piggybacking* (Figura 11). Em um sistema com somente *batching*, porcentagens inferiores ao valor alvo são obtidas para capacidade de sistema de 300 canais, enquanto que, em sistemas com *batching* e *piggybacking*, elas são atingidas para uma capacidade de sistema de 50 canais. Similarmente a sistemas com *batching*, em sistemas com *batching* e *piggybacking*, é pouco atrativo reservar um conjunto de contingência. Note que, em sistemas sem canais de contingência, tanto a probabilidade de abandono, quanto a porcentagem de operações de VCR rejeitadas, obtém-se valores alvos para capacidade menor que a necessária em sistemas com conjunto de contingência (Figura 10).

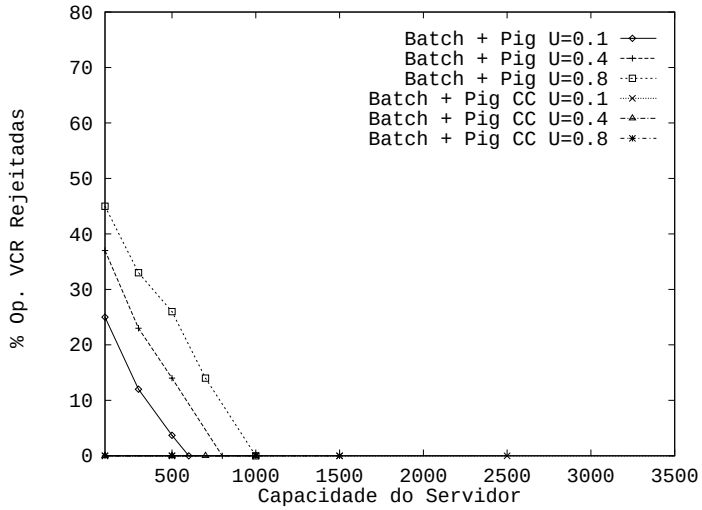

Figura 11: Porcentagem de op. VCR Rejeitadas X Capacidade do Servidor, taxa de 10 req/min.

As vantagens de se adotar *piggybacking* em conjunto com *batching* tornam-se evidentes ao se observar o sistema sob cargas altas. Enquanto que em um sistema com batching o número máxi-

mo de usuários admitidos no sistema é da ordem de 4500 (Figura 6), em um sistema com *batching* e *piggybacking* o número máximo de usuários admitidos chega a 8000 (Figura 12).

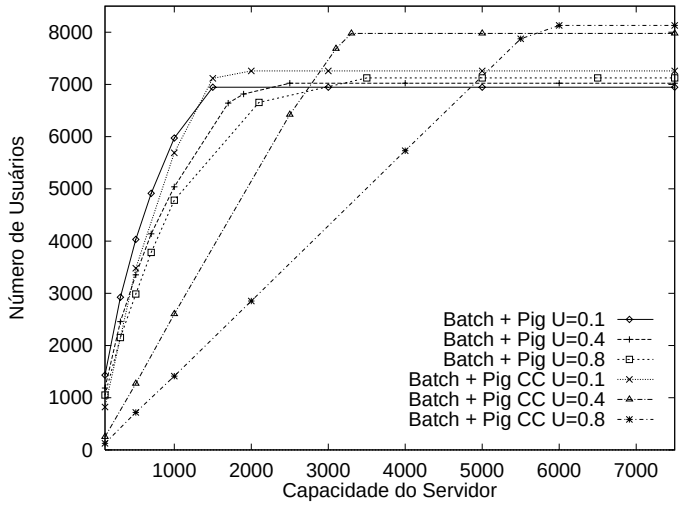

Figura 12: Número de Usuários Admitidos X Capacidade do Servidor, taxa de 60 req/min.

Sob altas cargas, obtem-se um valor para probabilidade de abandono de 1% com capacidade do sistema significativamente maior que em baixas cargas. Por exemplo, se, sob baixas cargas, para um grau de interatividade de 80%, em um sistema com canais contingentes, a demanda de canais é menor que 1500 , sob altas cargas (Figura 13) ela 'e de 5800 canais. A vantagem de se adotar conjuntamente *batching* e *piggybacking* pode, também, ser apreciada ao se comparar a menor demanda de canais para se obter o valor alvo de probabilidade de abandono em um sistema com *batching* e *piggybacking* (Figura 13) do que a demanda em sistemas com *batching* apenas(Figura 7).

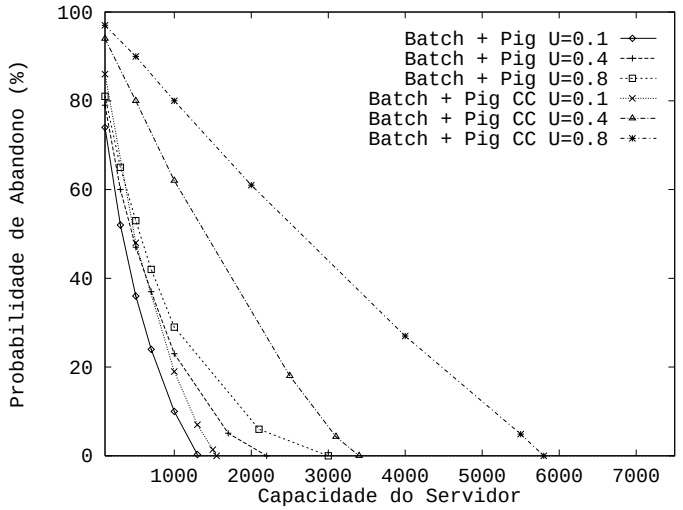

Figura 13: Probabilidade de Abandono X Capacidade do Servidor, taxa de 60 req/min.

Para um grau de interatividade baixo (10%), um sistema com canais de contingencia é preferível do que um sistema sem canais contingência, dado que os valores alvos da probabilidade de

abandono (Figura 13), bem como da percentagem de operacões VCR rejeitadas (Figura 14) são obtidas com aproximadamente o mesmo número de canais. No entanto, para valores médio e alto do grau de interatividade, os valores alvos são obtidos com um menor número número de canais, em um sistema sem canais de contingência. Por exemplo, para 80% de interatividade o conjunto das métricas a 1% é atingido com 3200 canais para sistemas sem canais de contingência, e com 5900 canais para sistemas com canais de contingência.

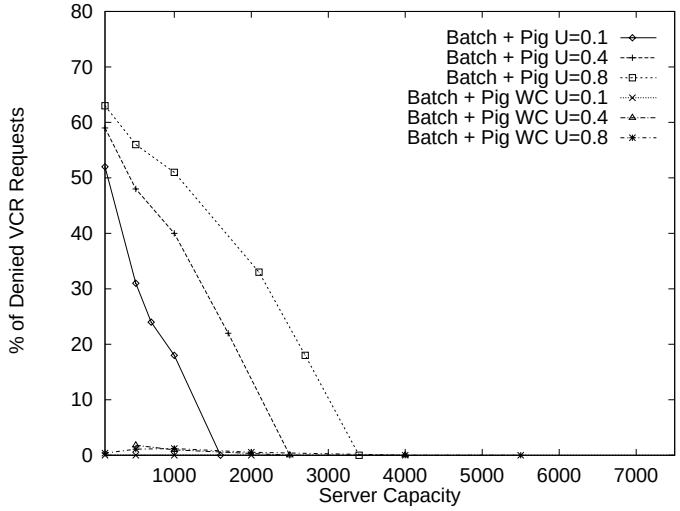

Figura 14: Porcentagem de op. VCR Rejeitadas X Capacidade do Servidor, taxa de 60 req/min.

É interessante notar que a suposição de que um fluxo de vídeo que retém um canal de contingência pode apenas se unir com o fluxo *multicast* original conduz a estimações conservadoras do número de canais demandados. Conseqüentemente, a conclusão de que sistemas sem reserva de canais de contingência são mais atrativos que sistemas com estes canais pode ser contestada. Entretanto, pode-se notar que os resultados aqui apresentados correspondem a um conjunto de experimentos com $N_{vcr} = 2$. Estimações baseadas em tal valor de N_{vcr} são próximas ao número médio de canais demandados, derivado via simulação (veja Figura 2), e, portanto, reforçam as conclusões apresentadas acima.

VII - Conclusões

Este artigo introduziu um modelo aproximado para determinar o número de canais necessários para dar suporte a operações de VCR em sistemas de VoD. O modelo é uma fila Erlang B cuja a taxa de chegada é uma aproximação da taxa de chegada de requisições de operações de VCR. A precisão da estimativa número de canais foi verificada por resultados de simulações. O número de canais estimado é um limite superior do número médio de canais demandados.

Além disso, o desempenho de diversos sistemas de VoD foi analisado. A eficiência de reservar canais de contingência para dar suporte a execução de operações de VCR foi investigada. Mostrou-se que, em sistemas sem canais de contingência, a meta de 1% para os valores da probabilidade de abandono, bem como da porcentagem de operações de VCR rejeitadas, é obtida com uma capacidade do sistema menor que em sistemas com o conjunto de canais de contingência. Além disso, um sistema sem conjunto de contingência aceita um maior número de usuários que um sistema com canais de contingência. Adicionalmente, mostrou-se que um sistema interativo com

batching e *piggybacking* admite um maior número de usuários, e provê menores probabilidade de rejeição de operações de VCR que um sistema com apenas *batching*.

Como trabalho futuro, sugere-se uma comparação entre sistemas em que requisições de operações de VCR são atrasadas quando não houver canais disponíveis, e sistemas nos quais requisições de VCR são rejeitadas ao invés de atrasadas.

Agradecimentos

Este trabalho foi parcialmente finaciado pelo CNPq e pela CAPES. Agradecemos aos Profs Edmundo Souza e Silva e Daniel Menascé pelas discussões sobre redes de filas.

Referências

[1] L. Golubchik, J. C. S. Lui e Muntz, "Adaptive Piggybacking: A Novel Technique for Data Sharing in Video-on-Demand Storage Servers", Multimedia Systems, 4(3):140--155, 1996.

[2] N.L.S. Fonseca e R. A. Façanha, "The Look-Ahead-Maximize-Batch Batching Policy", anais da conferência IEEE Global TelecommunicationsConference, pg 354-358, 1999.

[3] N. L. S. da Fonseca e R. A. Façanha, "Integrating Batching and Piggybacking in Video Server", anais da conferência IEEE GlobalTelecommunications Conference 2000, pg 1334-1338, 2000.

[4] P. Branch, C. Edgar e B. Sonkin, "Modeling Interactive Behavior of a Video Based Multimedia System", anais da conferência IEEE International Conference on Communications, 978-982, 1999.

[5] D. A. Menascé, V. A. F. Almeida e L. W. Dowdy, "Capacity Planning and Performance Modeling", PTR Prentice Hall, 1994.

[6] E. de Souza e Silva e R.R. Muntz, "Queueing Networks: Solutions and Applications", em Stochastic Analysis of Computer and Communication Systems, H. Takagi editor, North Holland, 1990

[7] M. Reiser, "Mean-Value-Analysis and Convolution Method for Queue-Dependent Servers in Closed Queueing Networks", Performance Evaluation, vol 1, pg 7-18, 1981.

[8] C. C. Aggarwal, J. Wolf e Philip S. Yu, "On Optimal Piggybacking Merging Policies for Video-on-Demand Systems", Proc. of the ACM Sigmetrics , vol 24, pp. 200--209, 1996

[9]R. A. Façanha e N.L.S. da Fonseca, "A Piggybacking Policy for Reducing the Bandwidth Demand of Video Servers", Managing QoS in Multimedia Networks and Services, J.N. de Souza e R. Boutaba (editores), pag 225-236,Kluwer Academic Publishers, 2000.

[10] N. L. S. da Fonseca e H. K. Rubinsztejn, "Channel Allocation in True Video-on-Demand Systems", anais da conferência IEEE GlobalTelecommunications Conference 2001, pg 1999-2004, 2001.

[11] N. L. S. da Fonseca e H. K. Rubinsztejn, "Dimensioning the Capacity of Interactive Video Server", anais da conferência International Teletraffic Congress 17, pg 383-394, 2001.