“Analisando o Impacto de Compartilhamentos no Dropbox em uma Rede Acadêmica”

Apresentação sobre o artigo científico

Índice

Resumo

Resumo

“Serviços de armazenamento na nuvem (e.g., Dropbox) são meios populares de compartilhamento de conteúdo e realização de trabalho colaborativo. Contudo, o compartilhamento nas nuvens pode levar a desperdício de largura de banda quando o mesmo conteúdo é recuperado de servidores remotos por múltiplos usuarios em um mesmo domínio de rede.

Este artigo apresenta uma caracterização dos padrões de compartilhamento no Dropbox a partir de dados de tráfego coletados de um campus universitário durante 4 meses. [...]”

Resumo

“[...] Em seguida, utilizamos os resultados da caracterização para desenvolver um gerador de cargas sinteticas que permite avaliar alterações no protocolo de sincronização do Dropbox. Propomos então uma arquitetura de sincronização que inclui caches para temporariamente armazenar as atualizações dos usuários.

Nossos resultados indicam que, mesmo com um cache pequeno, é possível evitar praticamente todos os downloads redundantes, o que beneficia provedores do serviço de armazenamento, usuarios e a Internet.”

Introdução

Introdução

Os serviços de armazenamento em nuvem vêm crescendo cada vez mais, e em ambientes acadêmicos o Dropbox ocupa grande fatia no leque de opções desses serviços, por isso foi escolhido como objeto de estudo.

Contudo, a utilização desse tipo de serviço causa impactos negativos em relação a consumo e desperdício de banda em uma rede.

Introdução

A aplicação possui mecanismos para otimizar o consumo da rede, como compressão, deduplicação e agrupamento de dados. O protocolo LAN Sync, *chave da aplicação*, faz com que em uma rede local de mesma subnet os arquivos em comum de usuários nessa rede não sejam atualizados da nuvem.

Contudo, isso não é o suficiente para otimizar de forma satisfatória uma rede, principalmente no caso das redes acadêmicas, onde devido a sua extensão, é comum a existência de diversas subnets, onde no caso a otimização não ocorre.

Introdução

Com base nessa deficiência do serviço, é feito um estudo para entendimento do funcionamento do serviço. Em seguida, é desenvolvido um gerador de cargas sintéticas para capturar e avaliar o funcionamento dos compartilhamentos no Dropbox.

Por fim, com base nesses dados, propor a mudança no funcionamento dos protocolos do serviço, a fim de criar um cache local na rede para armazenar as atualizações feitas pelos usuários para se evitar a redundância de downloads.

Trabalhos Relacionados

Trabalhos Relacionados

A caracterização feita neste trabalho e a construção do gerador de cargas sintéticas tomam como base resultados de [Drago et al. 2012], que apresenta **uma caracterização do tráfego do Dropbox a partir de medidas passivas** e de [Gonçalves et al. 2014], que **propõem um modelo para o funcionamento do cliente Dropbox**. Diferentemente desses trabalhos, o artigo apresentado:

- Caracteriza os padrões de compartilhamento no Dropbox;

- Desenvolve um gerador de cargas sintéticas;

- Propõe uma arquitetura de sincronização para auxiliar o serviço de armazenamento na nuvem.

Obs.: Outros trabalhos relacionados a esta temática são comentados na seção 2 deste artigo.

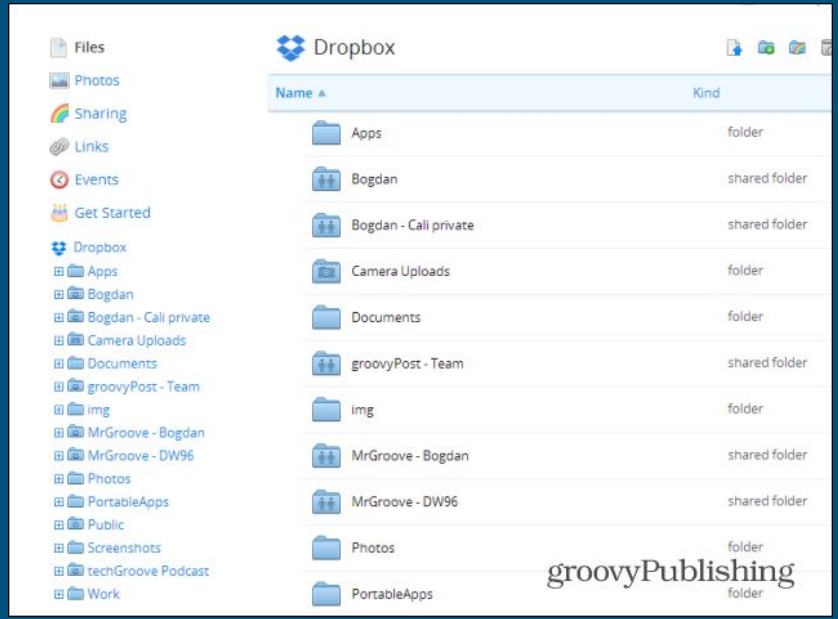
Conceitos e Metodologia

Conceitos e Metodologia - Dropbox: Visão Geral

Estrutura do Dropbox:

- Pasta Raíz do Usuário
  - Arquivos
  - Subpastas
  - Conteúdo compartilhado

Arquivos e subpastas podem ser compartilhados com outros usuários. As pastas compartilhadas são tratadas como pastas raízes, sendo exibido para o usuário assim que o convite é aceito.



Conceitos e Metodologia - Dropbox: Visão Geral

Componentes Dropbox:

- Servidores de Controle: Infraestrutura privada (contas, políticas, etc).
- Servidores de Armazenamento: Nuvem pública da Amazon (dados).

A troca de informações de autenticação e de dados é criptografada, contudo as informações de requisições de pasta, que ocorrem por minuto, são transmitidas por um servidor HTTP, logo não criptografado.

Conceitos e Metodologia - Coleta de Dados

A coleta de dados foi feita utilizando a ferramenta **tstat** em um campus universitário com cerca de 57 mil pessoas. Na coleta do tráfego de rede, foram capturados as mensagens que não são criptografadas, que armazenam as seguintes informações:

- IP mascarado do dispositivo conectado ao Dropbox.
- ID do dispositivo.
- ID associado ao usuário.
- ID único referente a cada pasta compartilhada no dispositivo.
- ID único e crescente para controle de versão das pastas.

Conceitos e Metodologia - Coleta de Dados

Um método para associar as notificações aos dados transferidos foi proposto observando que o cliente emite notificações logo após alterar uma pasta (upload) ou antes de receber uma modificação (download).

O tempo decorrido entre as transferências de dados e notificações mais próximas foi mensurado, ambos pertencentes ao mesmo endereço IP nos traços de dados. Foi adotado o limiar de 10 segundos para realizar a associação de notificações aos dados transferidos.

Conceitos e Metodologia - Coleta de Dados

Tabela 1. Sumário do tráfego Dropbox processado dos traços da rede.

Período	24/03–31/07/14
Volume Upload (TB)	1.97
Volume Download (TB)	4.39
Número de usuários	6478
Número de usuários que compartilham pastas	3445
Número de pastas	16485
Número de pastas compartilhadas localmente	3233
Número de notificações de modificação em pastas	3160095

Avaliação dos Compartilhamentos no Dropbox

Avaliação dos Compartilhamentos no Dropbox

- Limiares maiores que 10s aumentam a porcentagem de transferências associadas para todas as pastas;

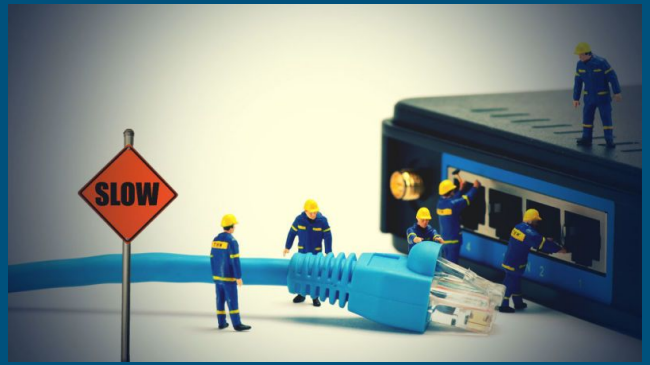
- O valor de limiar tem impacto mínimo na porcentagem de download para pastas compartilhadas, que se mantém em torno de 24% para o período da coleta;

Tabela 2. Estimativas dos volumes de *upload* e *download* (GB) para todas as pastas e para as pastas compartilhadas localmente para diferentes limiares.

	limiar de 10 seg.	limiar de 60 seg.	limiar de 600 seg.
Upload	871,07 (43%)	1201,88 (59,5%)	1518,08 (75%)
Download	942,12 (21%)	1333,17 (29,7%)	2021,59 (45%)
Upload (compartilhadas)	85,59 (9,8%)	117,26 (9,7%)	142,73 (9,4%)
Download (compartilhadas)	222,24 (23,6%)	326,23 (24,5%)	493,45(24,4%)

Avaliação dos Compartilhamentos no Dropbox

De acordo com os dados, percebe-se um desperdício de banda que não ocorreria se a questão de diversas subnets diferentes na mesma rede não fosse um problema.



Avaliação dos Compartilhamentos no Dropbox

Em uma outra analise de dados, é feito um estudo que visa entender os padrões de compartilhamento do Dropbox.

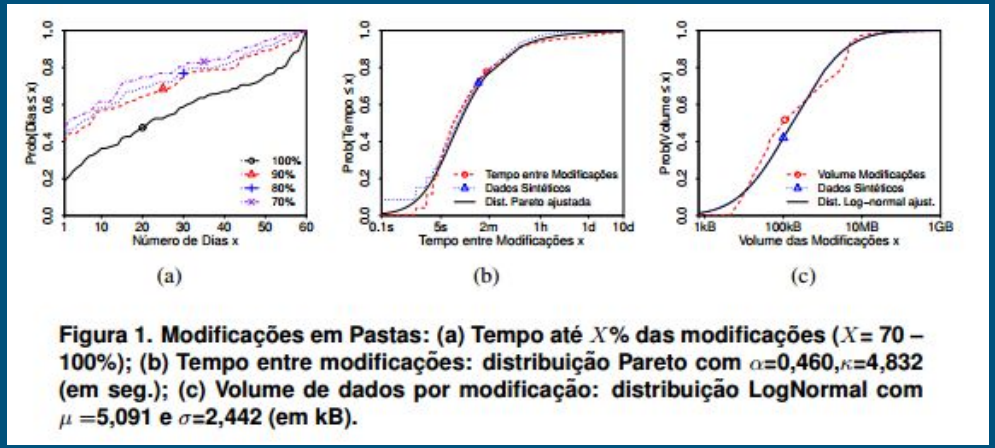
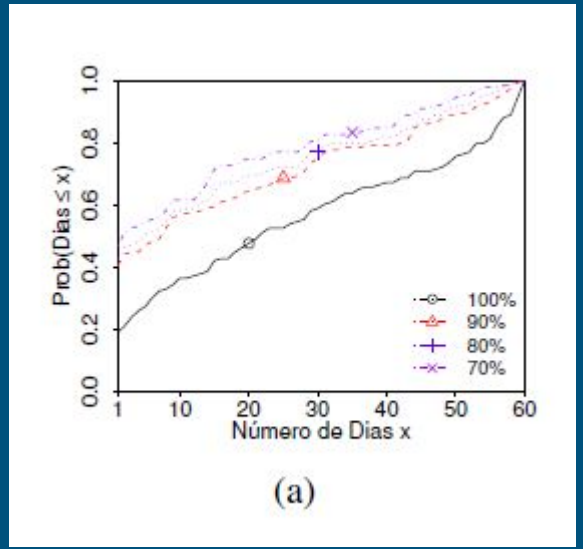


Figura 1. Modificações em Pastas: (a) Tempo até X% das modificações (X= 70 – 100%); (b) Tempo entre modificações: distribuição Pareto com α=0,460,κ=4,832 (em seg.); (c) Volume de dados por modificação: distribuição LogNormal com μ =5,091 e σ=2,442 (em kB).

Avaliação dos Compartilhamentos no Dropbox

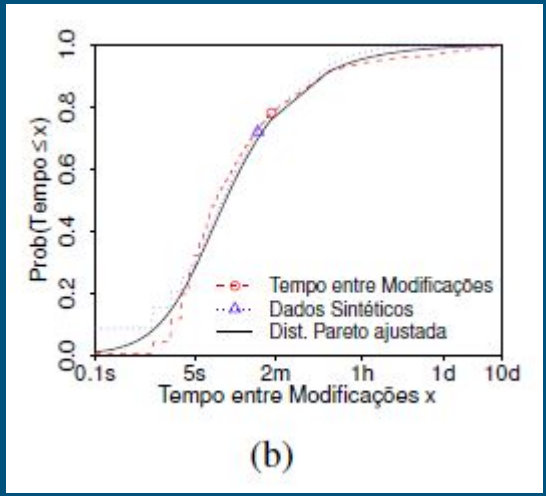
"A Figura 1(a) apresenta a distribuição acumulada do número de dias desde a criação de uma pasta até que a mesma tenha recebido X% de todas as modificações observadas em nossos dados, para X igual a 70%, 80%, 90% e 100%. [...]

Estes números indicam que muitas pastas recebem rajadas de modificações em curtos períodos de tempo. De fato, a figura mostra que cerca de 45% das pastas recebem 70% de todas as suas modificações apenas no primeiro dia de existencia."



Avaliação dos Compartilhamentos no Dropbox

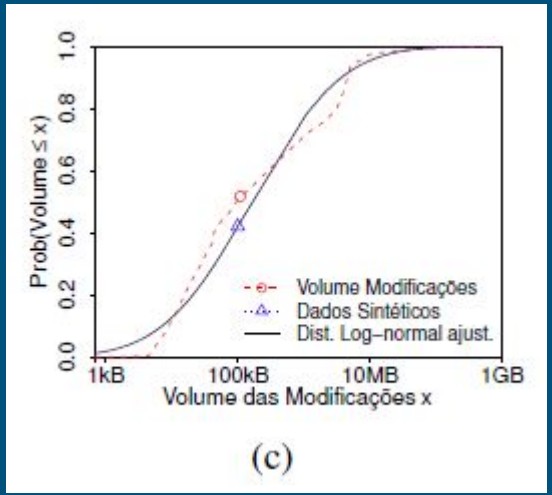
"A Figura 1(b) mostra a distribuicão acumulada computada sobre todos os intervalos entre modificações em uma mesma pasta, para todas as pastas. Mais de 70% das modificações ocorrem em intervalos inferiores a 1 minuto, embora alguns intervalos superem 1 dia."



Avaliação dos Compartilhamentos no Dropbox

“A última variável analisada é o volume de dados associado a cada modificação cuja distribuição acumulada, computada sobre todas as modificações de todas as pastas, é mostrada na Figura 1(c).

Observa-se que as modificações tendem a gerar um tráfego pequeno: cerca de 40% das modificações correspondem a menos de 100 kB, enquanto, na média, cada modificação corresponde a 3 MB.”



Gerador de Cargas Sintéticas

Gerador de Cargas Sintéticas

O gerador de cargas sintéticas tem por objetivo criar um traço de modificações em um dado número n de pastas do Dropbox compartilhadas pelos usuários durante um intervalo de tempo, visando capturar os volumes de dados transmitidos na rede ao longo do tempo à medida em que um usuário modifica uma pasta (upload) e os demais usuários locais que compartilham a mesma pasta obtêm essa modificação (download).

O funcionamento do gerador consiste na associação a cada pasta dos usuários (locais) que a compartilham.

O gerador faz diversas simulações utilizando a distribuição de duração de sessão e tempo entre sessões. Para cada pasta, também há uma simulação de uma sequência de modificações por usuários locais utilizando as distribuições dos tempos entre modificações e de volume de dados por modificação.

Gerador de Cargas Sintéticas

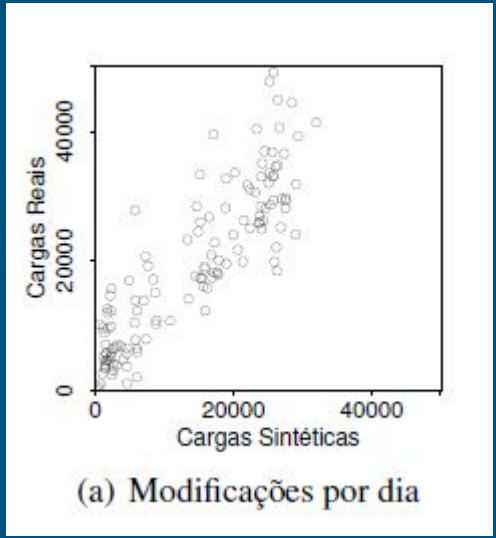
A validação do gerador ocorreu de duas maneiras:

- Avaliação que as distribuições das variáveis modeladas extraídas dos traços sintéticos se aproximaram das distribuições correspondentes aos dados reais;

- Número e o volume de modificações ocorridas diariamente nos traços coletados.

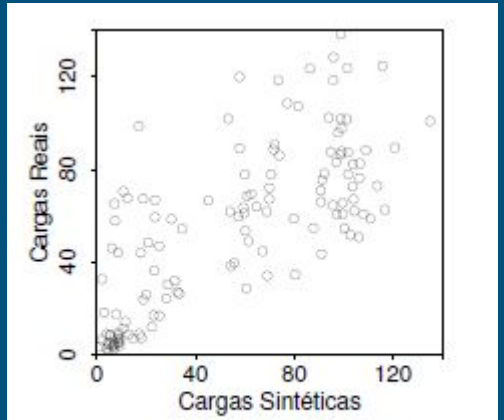
Validação do Gerador de Cargas Sintéticas

“A Figura 2(a) mostra que o gerador subestima o número de modificações, muitos dias possivelmente devido aos casos em que as modificações das pastas ocorrem quando usuários associados a ela não estão com sessões abertas.”



Validação do Gerador de cargas Sintéticas

“A Figura 2(b) mostra que o gerador tende a superestimar o volume gerado, o que é razoável uma vez que, para o planejamento de capacidade dos servidores de armazenamento ou da capacidade da rede, tais superestimativas levam a decisões mais conservadoras.”



(b) Volume (GB) por dia

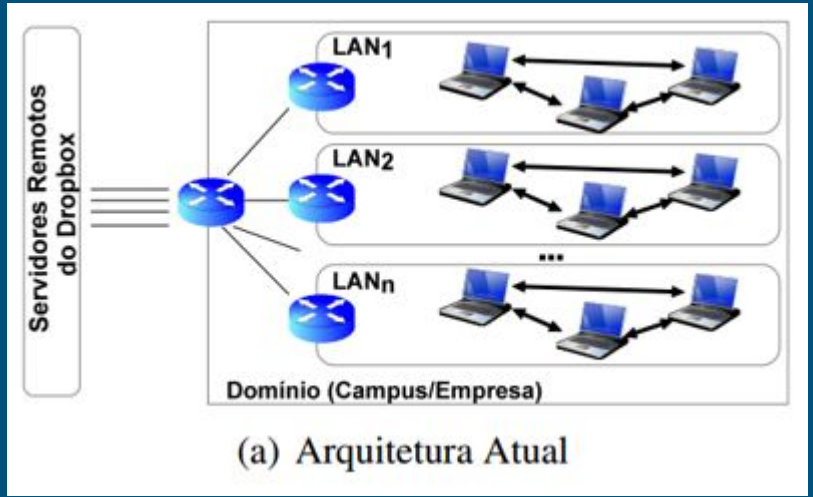
Nova Arquitetura de Sincronização

Descrição da Arquitetura Proposta

Arquitetura atual:

-Modificações são armazenadas em servidores remotos;

-Apresenta o uso de LAN Sync – permite que as modificações de pastas compartilhadas por usuários de uma mesma rede local sejam feitas diretamente entre eles.

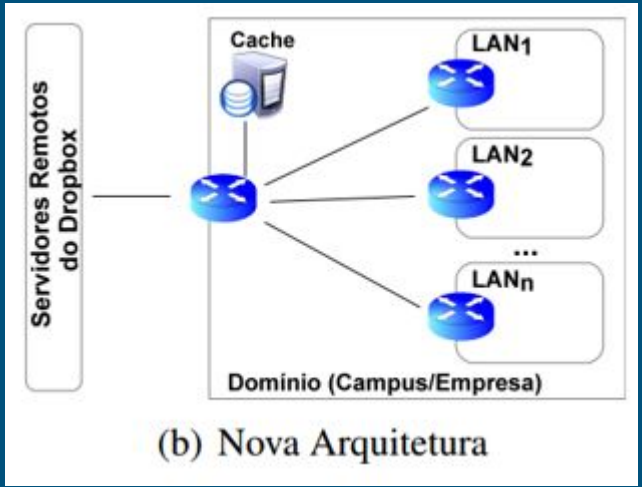


(a) Arquitetura Atual

Descrição da Arquitetura Proposta

Nova arquitetura:

- Considera a inclusão de um cache para armazenar dentro de cada domínio de rede as modificações de pastas de usuários do domínio.



(b) Nova Arquitetura

Descrição da Arquitetura Proposta

“Nossa proposta pressupõe que o Dropbox forneça o serviço especial de cache para universidades ou grandes empresas onde muitos usuários utilizam o cliente Dropbox. A universidade ou empresa interessada na implementação do serviço deve fornecer a infraestrutura necessária (i.e., um servidor) para a instalação e funcionamento do mesmo.”

Descrição da Arquitetura Proposta

“O serviço de cache funcionaria da seguinte forma. Qualquer modificação feita por um usuário locar é armazenada primeiramente no cache e posteriormente enviada aos servidores de armazenamento e controle do Dropbox. Em seguida, o Dropbox notifica essa modificação para todos os clientes de usuários que compartilham a pasta. Ao receber uma notificação, o cliente primeiramente busca a modificação no cache local. Se a modificação estiver armazenada no cache, ocorre um acerto (hit), caso contrário uma falta (miss). Na ocorrência de uma falta, o serviço de cache busca a modificação no Dropbox e a armazena no cache para servi-la ao cliente requisitante e a outras possíveis futuras requisições.”

Avaliação da Nova Arquitetura

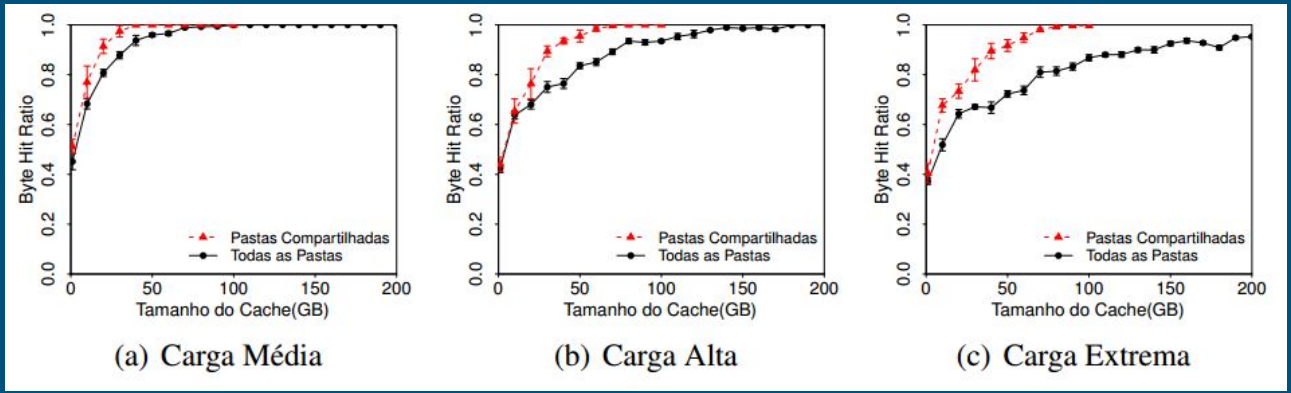
Os resultados dessa avaliação foram baseados nos dados coletados pelo gerador de carga sintética, que consideram tanto a redução do tráfego de download quanto o custo to cache.

Utiliza-se ***byte hit ratio*** = relação de bytes disponibilizados pelo cache em relação a quantidade de bytes requisitada por um ***client***.

Avaliação da Nova Arquitetura

A figura abaixo mostra o bit hit ratio em três cenários diferentes:

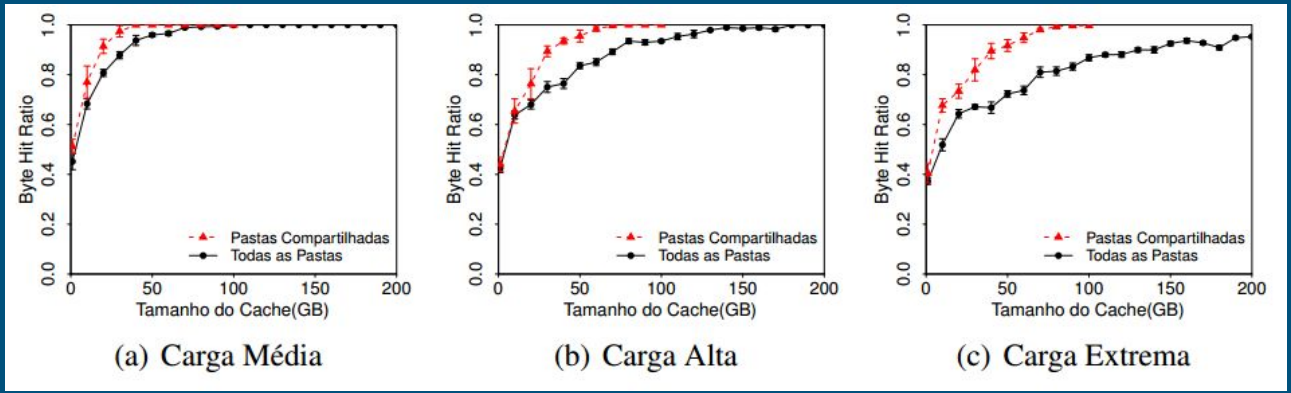
"(1) uma carga de 1065 pastas modificadas por dia, o que corresponde a carga média observada em nossos traços, (2) uma carga de 2185 pastas, que representa o maior numero de pastas observadas diariamente em nossos traços, e (3) uma carga bem mais alta com 4000 pastas."



Avaliação da Nova Arquitetura

“A curva preta contínua mostra os resultados quando o cache armazena todas as modificações de pastas feitas, independentemente do número de usuários locais que compartilham a pasta.

A curva vermelha tracejada mostra os resultados quando o cache armazena apenas modificações de pastas compartilhadas por múltiplos usuários locais.”



Avaliação da Nova Arquitetura

Seguindo dos resultados, observa-se que:

- Um cache com apenas 1GB leva a um byte hit ratio de 40% nos 3 cenários.
- Um cache com 100GB resulta em um byte hit ratio superior a 93% nos cenários de carga média e alta.
- Um cache com 100GB resulta em um byte hit ratio igual a 87% no cenário de carga extrema.
- Um cache com mais de 150GB traz um aumento de byte hit ratio próximo de zero, trazendo poucos benefícios.

* *Lembrando:* Esta avaliação considerou tanto a redução no tráfego de download de modificações de pastas compartilhadas por usuários do domínio alvo, quanto o custo do cache.

Avaliação da Nova Arquitetura

Numa segunda avaliação, onde o cache armazena somente modificações de pastas compartilhadas por múltiplos usuários locais, é ainda mais eficiente:

- Um cache com 50GB leva a um byte hit ratio superior a 95% nos cenários de carga média e alta.
- Um cache com 50GB leva a um byte hit ratio igual a 90% no cenário de carga extrema.

* Contudo, esse segundo modelo possui uma complexidade de implementação maior, uma vez que as modificações de pastas devem ser tratadas de forma diferente, dependendo se a pasta é compartilhada localmente ou não.

Conclusão e Análise

Conclusão

Referente aos estudos realizados e expostos nesse artigo científico, concluímos que:

- O artigo aponta um problema que merece atenção por ser de um serviço popular em diversos ambientes, bem como no qual foi utilizado como objeto de estudo.
- Os dados coletados e os números que foram apresentados reforçam a atenção sobre o desperdício de banda causado pelo uso do serviço.
- Mesmo o Dropbox possuindo um protocolo para amenizar o desperdício de banda, ele não é suficiente.

Conclusão

Sobre o estudo como um todo e aos resultados obtidos, vemos como pontos positivos:

- A existência de estudos anteriores.
- A preocupação de se manter uma rede estável.
- O fator da escolha do software/serviço alvo por ser utilizado em grande escala.
- A exposição do entendimento e funcionamento do software/serviço.
- Resultados obtidos por captura de dados que reforçam a exatidão de infos.
- Resultados obtidos com o uso do cache que transparecem sua eficácia.

Conclusão

Sobre os pontos negativos do artigo e do estudo embasado:

- O software/serviço é privado e de código fechado, o que não permite alterações diretas no mesmo.
- Os números são representados por um gerador de carga sintética e não por uma aplicabilidade real de uma nova arquitetura.
- Uma possível modificação no software/serviço, em sua arquitetura, desencadearia esse estudo podendo torná-lo não utilizável.

Fonte:

http://sbrc2015.ufes.br/wp-content/uploads/proceedingsSBRC2015.pdf

Obrigado!

Dúvidas?

Cainã Cesar de Godoy
Simone de Cassia Santos