

Identificação e cache de aplicativos web dinâmicos: Uma abordagem flexível na solução de problemas de desempenho

Visão Geral A implementação de aplicativos web dinâmicos na companhia oferece benefícios importantes, mas os usuários também podem experimentar um desempenho imprevisível do aplicativo. Uma das causas do mau desempenho é a sobrecarga do servidor, resultado do maior uso do aplicativo e da complexidade do conteúdo.

O WebAccelerator da F5 resolve esse problema usando sua tecnologia Dynamic Caching (caching dinâmico). Essa é uma capacidade única, que ajuda a distribuir aplicativos web altamente dinâmicos aos usuários, de 5 a 10 vezes mais rápido, reduzindo muito os custos da infra-estrutura web. Essa função é particularmente relevante para aplicativos chegando ao limite da escalabilidade e com necessidade de suportar mais usuários, bem como aplicativos web que trabalham de forma lenta por causa de design legado e outros problemas.

Desafio O desafio principal gira em torno do aumento do desempenho do usuário final e do alívio da carga no servidor para os aplicativos e do conteúdo dinâmico da web. De maneira geral, o caching estático só pode ser usado para 30% dos pedidos HTTP, uma porcentagem que normalmente não inclui dados dinâmicos de alto valor (resultados de pesquisas, dados XML e assim por diante).

Solução

Caching dinâmico

O caching dinâmico muda completamente o modelo de caching, tornando possível usar essa tecnologia com uma variedade muito maior de conteúdo, incluindo páginas web dinâmicas, resultados de pesquisas e objetos XML. Essa tecnologia patenteada é exclusiva da F5 e não está disponível em nenhum outro fornecedor.

O caching dinâmico é focado na lógica e no comportamento do aplicativo, não apenas em objetos web individuais. Ao entender a lógica de alto nível dos aplicativos (o que pode ou não ser colocado no cache, quais eventos invalidam os dados, etc.), o WebAccelerator elimina o processamento repetitivo de pedidos web complexos. O caching dinâmico permite que o WebAccelerator decida quando invalidar objetos e como identificar partes reutilizáveis de conteúdo. Isso é possível graças às políticas de aceleração predefinidas do aplicativo, uma interface de usuário intuitiva, uma poderosa API baseada em XML (ESI) e uma funcionalidade de ativação baseada em solicitações HTTP que, juntas, oferecem controles completos para a validação e invalidação de conteúdo.

Sem o WebAccelerator e o caching dinâmico, as soluções existentes de caching possuem apenas a data de expiração do objeto como referência. O caching dinâmico permite que o cache verifique qualquer coisa em uma solicitação HTTP – de URLs a cookies, parâmetros de pesquisa e outros cabeçalhos – e produza invalidações inteligentes e chaves de cache. Além disso, o caching dinâmico permite que o WebAccelerator decida quando invalidar objetos e como identificar partes reutilizáveis de conteúdo.

Ao utilizar o Dynamic Caching, o WebAccelerator pode responder diretamente a até 80% das solicitações de usuário com maior carga computacional, sem envolver o resto da infra-estrutura do site. Além disso, o WebAccelerator não se confunde com a semântica dos aplicativos e nunca envia itens inválidos do cache.

Caching estático

Como uma extensão da sua capacidade de caching dinâmico, o WebAccelerator também oferece caching estático. O caching estático simplesmente serve objetos – normalmente imagens, javascript, folhas de estilo –, desde que eles não tenham ultrapassado sua data de expiração. Ainda que o caching estático provavelmente já exista na infra-estrutura computacional de um aplicativo, permitir que o WebAccelerator sirva objetos estáticos remove a carga de outra operação do aplicativo original.

Identificando o conteúdo

Para fazer o cache de aplicativos, é necessário identificar com precisão as partes individuais do conteúdo. Na primeira vez em que o WebAccelerator recebe uma solicitação para uma determinada parte de conteúdo, ele envia a solicitação para os servidores originais e recolhe aquele conteúdo do usuário. Quando o WebAccelerator recebe a página – antes de enviar a resposta ao cliente –, ele transforma uma cópia da página em uma representação interna compilada. Ele então usa essas respostas compiladas para recriar uma página, mediante uma solicitação HTTP.

As respostas compiladas são uma representação interna da página, conforme recebida do site original, bem como instruções descrevendo como recriar a página a partir da representação interna, incluindo informações necessárias para atualizar a página com qualquer conteúdo aleatório ou rotativo. O WebAccelerator também define um identificador de conteúdo exclusivo (Unique Content Identifier – UCI) para a resposta compilada, com base nos elementos presentes na solicitação, como a URI, parâmetros de pesquisa, etc. Essa informação é armazenada no cache como uma resposta compilada sob o UCI, que é usada tanto para a solicitação quanto para a resposta compilada criada para atendê-la. Quando uma solicitação futura for recebida e gerar os mesmos elementos e a mesma UCI, essa resposta é encontrada no cache e usada para atender à solicitação.

O uso desses elementos como parte do identificador facilita a combinação das solicitações futuras com o conteúdo correto no cache. Os elementos da solicitação HTTP que não afetam o conteúdo da página são ignorados e não são usados na UCI. Como resultado, os valores definidos para esses elementos não são usados para identificar instâncias exclusivas do conteúdo do cache.

Mas nem todos os elementos de uma solicitação indicam uma resposta exclusiva. Por exemplo, duas solicitações com a mesma URI, mesmo método e mesmos parâmetros de pesquisa, mas com cookies diferentes, podem gerar respostas idênticas dos servidores originais. Por padrão, o WebAccelerator assume que certos elementos podem causar a variação do conteúdo, e outros, não. Essa informação também é usada para criar a UCI e pode ser configurada usando as políticas de variação do WebAccelerator.

Exemplos

Imagine um aplicativo de helpdesk que gerencia um grupo de ordens de serviço em constante mudança. Embora cada ordem possa ser referenciada de várias formas, cada uma representa um objeto individual a ser colocado no cache. Para identificar as ordens individuais, o aplicativo depende de identificadores exclusivos gerados por sua base de dados. Usuários diferentes podem ver as mesmas ordens. As solicitações terão o seguinte formato:

```
GET /helpdeskapp/viewticket.asp?ticketid=121& parentname>Login HTTP/1.1  
Host: helpdesk.companhia.com Cookie: userid=323; sessionid=3xx3s
```

Aqui, o WebAccelerator seria configurado para identificar e armazenar a ordem de serviço usando esta chave:

[\[/helpdeskapp/viewticket.asp, ticketid=121\]](#).

Agora, vamos comparar isso com outro aplicativo, como um cliente de e-mail que depende principalmente da identidade do usuário para gerar conteúdo. Ao fazer o caching da caixa de entrada, as solicitações são feitas dessa forma:

GET /mymail/inbox.asp?page=1& parentname>Login HTTP/1.1 Host: meucorreio.companhia.com Cookie: userid=323; sessionid=3xx3s

O WebAccelerator, então, pode ser configurado para identificar esse objeto usando a seguinte chave:

[\[/mymail/inbox.asp, page=1, userid=323\]](#).

Para um caching eficaz do conteúdo dinâmico, é necessário manter a flexibilidade na identificação do conteúdo. Para manter a funcionalidade, o sistema deve sempre armazenar partes diferentes do conteúdo no cache como objetos separados, e deve haver relacionamentos um-para-um entre objetos verdadeiros e chaves de cache.

Mantendo a fidelidade de aplicativos

O conteúdo gerado por uma aplicação web muda constantemente. Os eventos que conduzem às mudanças pertencem a três categorias gerais: tempo, eventos de usuários e eventos de aplicativos. Em cada um deles, o WebAccelerator possui mecanismos no caching dinâmico que podem ser usados para descrever quando as mudanças ocorrem e quais objetos no cache são afetados. O WebAccelerator suporta a invalidação do cache baseando-se em:

- **Data de expiração**, útil para o caching de conteúdo estático ou quase estático. Como exemplo, um aplicativo gerando relatórios financeiros semanais. Os relatórios anteriores podem ser armazenados indefinidamente, e o relatório com o título "atual" é válido por uma semana. O WebAccelerator invalida o objeto no cache assim que ele expira.
- **Eventos de usuário**, causados pela interação do usuário com o aplicativo. O WebAccelerator monitora solicitações HTTP que combinam com certos critérios conhecidos de mudança de estado do aplicativo. Quando o WebAccelerator localiza uma solicitação assim, executa uma regra para invalidar a seção relevante do cache. Os fóruns são um bom exemplo. Os tópicos individuais são lidos com muito mais frequência do que escritos, portanto o caching oferece grandes vantagens. Para implementar o caching dinâmico, o WebAccelerator é configurado para reconhecer quando um usuário está escrevendo em um tópico específico (veja a Figura 1, abaixo). Ele armazena o texto do tópico, obtendo-o da primeira vez diretamente do aplicativo e, subsequente, do cache. O WebAccelerator reconhece quando alguém escreve em um tópico e invalida os objetos do cache associados somente àquele tópico. Quando o próximo usuário solicitar o tópico, o WebAccelerator o recuperará do aplicativo, semeando o cache novamente.

Figura 1



Os eventos de usuários invalidam objetos individuais no cache do WebAccelerator da F5

- **Eventos de aplicativos**, originados fora da interação entre usuário e aplicativos. O WebAccelerator gerencia os eventos de aplicativos aceitando mensagens de invalidação XML padronizadas (ESI) quando um evento de aplicativo ocorre.

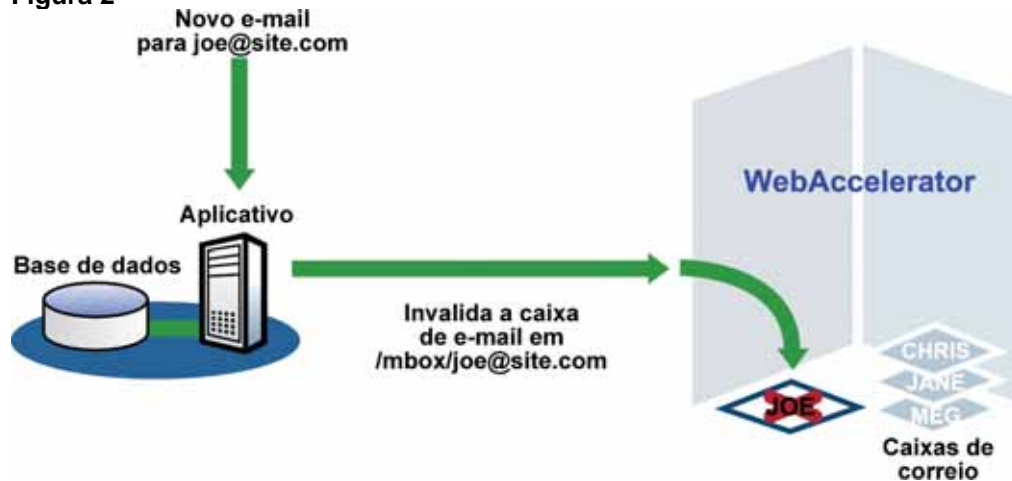
Exemplo de mensagem XML de invalidação de cache

A mensagem XML de invalidação de cache abaixo invalida todos os objetos gerados por `/myapp/somepage.jsp` que foram criados usando os cookies `plano=gold` e `grupo=ADMINISTRADOR`.

```
<?xml version="1.0"?>
  <!DOCTYPE INVALIDATION SYSTEM "invalidation.dtd">
<INVALIDATION VERSION="WCS-1.0">
  <OBJECT>
    <ADVANCEDSELECTOR URIPREFIX="/myapp/somepage.jsp">
      <COOKIE NAME="plano" VALUE="gold"/>
      <COOKIE NAME="grupo" VALUE="ADMINISTRADOR"/>
    </ADVANCEDSELECTOR>
  </OBJECT>
</INVALIDATION>
```

O recebimento de uma nova mensagem é o exemplo de um evento externo para um aplicativo de e-mail (veja a Figura 2, abaixo). Quando não tiver novos e-mails, o WebAccelerator fornece a resposta diretamente do cache. Quando chega uma nova mensagem, uma notificação é enviada do aplicativo de e-mail para o WebAccelerator, indicando mudanças na caixa de entrada de um determinado usuário. Então, quando o usuário acessar novamente sua caixa de entrada, o WebAccelerator encaminha à solicitação ao aplicativo de e-mail, atualizando a lista da caixa de entrada.

Figura 2



Os eventos de usuários invalidam objetos individuais no cache do WebAccelerator

Outro exemplo é a mudança de preços de produtos em um site comercial. As mudanças de preço acontecem constantemente e cerca de 1% das páginas de um site comercial muda todos os dias. Quando ocorre uma mudança de preços, o WebAccelerator exclui a página armazenada anteriormente no cache. Isso não é possível com o caching estático de outros fornecedores.

Conclusão

O WebAccelerator resolveu o problema persistente do cache de conteúdo dinâmico implementando duas capacidades fundamentais:

- Um sofisticado algoritmo de combinação, que vincula pesquisas qualificadas de usuários ao conteúdo do cache
- Um mecanismo de invalidação de cache disparado por eventos de usuários e aplicativos

O caching dinâmico reduz a carga do servidor e a latência em até 80%, melhora o desempenho do usuário, reduz os custos de infra-estrutura, oferece capacidade de substituição durante interrupções do site (planejadas ou não) e controle absoluto sobre a precisão do conteúdo, sem mudanças na arquitetura ou no código do site. Além disso, ele oferece integração perfeita com outros aplicativos e suporta uma grande variedade de software e tipos de site, independentemente de elementos específicos da infra-estrutura.

Sobre a F5

A F5 Networks é a líder global em Application Delivery Networks. A F5 fornece soluções que tornam os aplicativos seguros, rápidos e disponíveis para todos, ajudando as companhias a obter o maior retorno pelo seu investimento. Ao implementar inteligência e gerenciabilidade na rede para transferir a carga de aplicativos, a F5 os otimiza, permitindo que eles trabalhem mais rápido e consomem menos recursos. A arquitetura expansível da F5 integra de forma inteligente a otimização de aplicativos, protege os aplicativos e a rede e oferece confiabilidade aos aplicativos - tudo em uma plataforma universal. Mais de 10.000 companhias e provedores de serviços em todo o mundo confiam na F5 para manter seus aplicativos funcionando. A companhia tem sede em Seattle, Washington, com escritórios no mundo todo. Para mais informações, visite www.f5.com (em inglês).