

Benefícios da Otimização de Aplicativos BIG-IP na WAN

Visão Geral Aplicativos Web se tornaram comuns nas empresas de hoje. Não é incomum que uma única empresa tenha centenas de aplicativos Web em sua rede. À medida que implementar esses aplicativos se torna mais fácil, o foco está se deslocando dos aplicativos para as preocupações de segurança e acesso. Um dos principais desafios com relação aos aplicativos Web é tentar acompanhar o aumento exponencial de usuários que acessam esses aplicativos remotamente. Não só isso apresenta novas questões de segurança, como também os usuários remotos exigem o mesmo nível de desempenho que experimentam ao se conectar aos aplicativos através da rede local. Resolver questões de desempenho e segurança com a personalização de cada aplicativo não é apenas extremamente caro, como também é um consumo ineficiente de tempo e recursos. Com sua posição estratégica na infra-estrutura de rede, o novo sistema BIG-IP oferece uma solução rápida, fácil e muito mais econômica a esses desafios. A versão 9.x do sistema BIG-IP inclui alguns recursos que facilitam a otimização e aceleração do tráfego de aplicativos. O F5 Networks Solution Center realizou extensos testes para medir o ganho de desempenho que o sistema BIG-IP dá aos aplicativos e websites acessados pela Internet.

Esses testes demonstram as capacidades únicas da tecnologia de aceleração de aplicativos do dispositivo BIG-IP, o que o diferencia no mercado. Embora outros fornecedores afirmem a capacidade de seus produtos de melhorar o desempenho de aplicativos, seus testes foram feitos num ambiente de laboratório controlado. Isso deixa aos usuários a difícil tarefa de tentar determinar como esses resultados se traduzem em ganhos de desempenho em situações reais. É a essa pergunta que a F5 se dispôs a responder. E chegamos à conclusão de que a única maneira de testar o desempenho real para usuários pela Internet era realmente usar a Internet.

Para executar esses tipos de testes reais, a F5 escolheu a Gomez® Performance Network (descrita em detalhes na seção a seguir) para testar transações ponto a ponto. Em média, o sistema BIG-IP melhorou os tempos de resposta para usuários finais, os reduzindo à metade ou mais; reduziu a utilização de largura de banda em mais de **75%**; reduziu a ocorrência de erros de limite de tempo em navegadores por links lentos em mais de **80%** e reduziu a carga de até **98%** das conexões para os servidores.

Solução

Acelerando os Aplicativos e os Tempos de Resposta para Usuários Finais

O dispositivo BIG-IP contém otimizações direcionadas e especializadas que se aproveitam das exclusivas tecnologias da F5 para aceleração de WAN, LAN e dados. Isso permite que o sistema BIG-IP ofereça uma otimização inigualável, recuperação de perda de pacotes e uma mediação mais inteligente entre servidores não-ótimos e clientes. A Tabela 1 mostra a melhoria nos tempos de resposta para alguns aplicativos típicos.

Aplicativo	Não Otimizado	Otimizado com BIG-IP	Melhoria
BEA WebLogic Portal 8.1	38,209 segundos	17,291 segundos	121% (2,2x)
Microsoft IIS 6.0	21,09 segundos	12,39 segundos	70% (1,7x)
Microsoft Outlook Web Access 2003	32,88 segundos	21,39 segundos	55% (1,7x)
Microsoft SharePoint Portal Services 2003	40,6 segundos	18,06 segundos	125% (2,2x)
Siebel Business Applications 7.7	88,157 segundos	33,989 segundos	126% (2,3x)

Tabela 1: Amostras de Melhorias para Transações Completas de Aplicativos

Melhorando a Capacidade do Grupo de Servidores

Além de oferecer uma melhor experiência para os usuários, também está demonstrado que o dispositivo BIG-IP melhora significativamente a escalabilidade da infra-estrutura ao redor.

Usando o novo recurso **Fast Cache**, o sistema BIG-IP reduziu a carga de, em média, **36%** das tarefas de entrega de conteúdo e **95%** das conexões TCP dos servidores de back-end, oferecendo ganhos drásticos de desempenho e escalabilidade para a infra-estrutura existente.

Por meio dessas e de outras técnicas de otimização, comprovamos que, com o sistema BIG-IP na rede, os servidores geralmente se tornam capazes de suportar o dobro da carga de trabalho.

Aumentando a Eficiência da Largura de Banda

Sem otimizações, as empresas geralmente só conseguem aproveitar uma parte da largura de banda que adquiriram, por causa de ineficiências de WAN e de protocolos subjacentes. Através do TMOS (Traffic Management Operating System – Sistema Operacional de Gerenciamento de Tráfego) e do conjunto de recursos TCP Express, os testes revelaram que o dispositivo BIG-IP melhora consideravelmente a eficiência de largura de banda para um site. Por exemplo, usando o sistema BIG-IP, testes mostraram um **aumento médio de 224% dos dados passados pelo cabo (3,2 vezes)** e uma **redução média de 50% dos pacotes** que trafegam na rede.



Figura 1: Amostras de redução da banda usando compressão inteligente e otimizações TCP

De forma geral, o sistema BIG-IP oferece **um aumento médio de 322% (4 vezes) na eficiência de utilização de largura de banda.**

Melhorando a Confiabilidade das Conexões WAN

Usando o conjunto de recursos TCP Express, testes demonstraram que o sistema BIG-IP reduziu em até 50% o número de erros de limite de tempo de TCP e de conexão de TCP restaurada vistos pelos clientes. Isso é especialmente benéfico para transmissões por redes de alta perda, ou para clientes usando conexões de baixa largura de banda, como conexões discadas.

Testes no Mundo Real Usando a Gomez Performance Network

Quando a F5 decidiu testar o desempenho de nossas capacidades de otimização de aplicativos de maneira a oferecer valor aos nossos clientes, foi necessário encontrar um serviço terceirizado confiável que pudesse nos dar acesso a uma população global e diversa de usuários, com vários graus de largura de banda e ambientes operacionais diferentes. Depois de analisar as soluções disponíveis, determinamos que a Gomez Performance Network (GPN) era nossa melhor opção.

A premiada Gomez Performance Network é formada por mais de 10.000 computadores em todo o mundo, os quais utilizam conexões reais de Internet para usuários finais. Essa ferramenta é usada por muitos de nossos clientes corporativos e já ganhou inúmeros prêmios, inclusive o de Escolha do Editor da Network Computing. Suas redes foram a ferramenta de teste que ofereceu à F5 os meios de coletar uma amostra correta e realista dos ganhos de desempenho para usuários finais na Internet.

Usando a GPN, os resultados dos testes do F5 Solution Center são muito mais realistas do que os materiais de marketing disponibilizados no mercado pela concorrência. É muito fácil criar testes que mostram uma melhoria de dez, vinte ou mais vezes, selecionando cuidadosamente as páginas testadas e os dados usados em uma aplicação dinâmica em ambiente de laboratório. Mas de que servem esses números se eles não representam um cenário passível de ser encontrado pela maioria dos usuários? Concluímos que esses resultados gerados em laboratório não são o que nossos clientes atuais e potenciais considerariam úteis ao tentar escolher um produto de aceleração de aplicativos – as pessoas querem ver qual melhoria de desempenho seus usuários finais podem conseguir.

A F5 decidiu que a melhor maneira de demonstrar o desempenho melhorado para aplicativos seria escolher aqueles que os usuários costumam ver sendo operados pela WAN. Escolhemos aplicativos de portal e de colaboração, como o BEA® WebLogic®, Siebel® Business Applications e Microsoft® SharePoint®. Também incluímos o Microsoft Outlook® Web Access como um aplicativo essencial que os usuários utilizam pela WAN, e um site genérico (www.f5.com) hospedado com o Microsoft Internet Information Services (IIS) para que se tenha uma idéia dos ganhos de desempenho em geral que podem ser obtidos para a maior parte do conteúdo estático. Embora as empresas possam estar executando uma multidão de aplicativos, os resultados vistos aqui são bons indicadores gerais dos ganhos de aplicações que podem ser obtidos usando esta tecnologia.

Para cada teste, medimos o tempo de resposta da experiência completa de um usuário final com um aplicativo, tirando-se a média com vários dias e milhares de

testes repetidos. Por exemplo, o script que empregamos para testar o Microsoft Outlook Web Access incluía o login, recuperação de mensagens, exibição de calendário, e assim por diante. Como vocês podem ver, isso é completamente diferente de tempos de download de uma única página, e representa o tempo de resposta de aplicativo ponta-a-ponta para uma transação completa, como um usuário realmente faria.

Metodologia dos Testes

A Gomez oferece um serviço que permite aos clientes monitorarem o desempenho de um website da forma vista pelos agentes de monitoramento da Gomez, executados em computadores de usuários reais por meio da Internet e por todo o mundo. A Gomez paga a empresas e usuários para emprestarem seus recursos de computadores e largura de banda. A Gomez, por sua vez, instala um agente no computador que o instrui a se conectar a determinados sites. Em intervalos regulares, a Gomez instrui um certo número de computadores a solicitar um website em particular e medir cada aspecto do seu desempenho. Essas informações são enviadas para a Gomez Performance Network para que o cliente possa ver a qualidade do desempenho do website para o usuário final.

Além de escolher o website solicitado, a Gomez permite aos clientes escolher qual população de usuários (ou "população de pontos") é usada, e oferece a capacidade de especificar o tipo de conexões de clientes que eles esperam servir, para a coleta de métricas de desempenho. Por exemplo, a F5 selecionou uma variedade de regiões de origem dos usuários, e a categoria de largura de banda das quais eles fazem parte. A Gomez oferece um registrador para facilitar a configuração dos exemplos de transações com usuários usadas pelos clientes, o que permite aos clientes especificar quais links são selecionados em cada website, quais credenciais de login são usadas (se aplicável) e outras configurações de transação necessárias para simular uma experiência real de usuário. Uma vez que o exemplo de sessão de usuário for configurada, usando a ferramenta de registro oferecida, essa sessão é reproduzida pelos agentes escolhidos na Gomez Performance Network. Ao testar o tempo de resposta de aplicativos para usuários finais reais pela Internet, descobrimos que estes são os aspectos mais importantes da Gomez Performance Network:

- Testes de última milha – esses testes usam usuários reais em todo o mundo.
- Definir as categorias de largura de banda que representam o público-alvo e a região-alvo.
- As principais métricas, o tempo de resposta "ponto-a-ponto" e os erros – as outras estatísticas ajudaram a detectar outros problemas.
- Registrar exemplos de sessão de usuário em vez de fazer downloads de páginas simples – capturar uma sessão de usuário típica em vez de fazer o download de uma única página.
- Calcular a média com dados resultantes de alguns dias para determiná-la com precisão, e trabalhar para melhorar essa média, modificando o aplicativo ou adquirindo tecnologias de melhoria de desempenho.
- Os agentes da Gomez suportam todos os comportamentos e recursos comuns dos navegadores, mas são limitados em sua capacidade para aceitar plug-ins de terceiros, assim como muitos usuários finais - o aplicativo deve ser capaz de lidar com esse tipo de cliente seguro e que não necessariamente aceita plug-ins.

Nossos parâmetros de teste foram bem simples; escolhemos usuários de conexão discada global, conexão discada nos EUA e banda larga menor nos EUA. Para cada aplicativo, usamos um exemplo de sessão de usuário e que envolvia etapas semelhantes às seguintes.

1. Ir para a página principal.
2. Clicar no botão para digitar as credenciais de login.
3. Digitar as credenciais de login.
4. Clicar no link A.
5. Clicar no link B.

Esse é apenas um exemplo em alto nível; na prática, as etapas variaram para atender a uma experiência de usuário típica para cada aplicativo.

Uma vez que a Gomez tinha nossa população de usuários e que tínhamos registrado uma amostra de experiência de usuário, usamos essa amostra para solicitar métricas de desempenho para duas instâncias de aplicativos simultaneamente. Uma instância de aplicativo foi acelerada com o sistema BIG-IP, e a outra foi executada com um simples balanceamento de carga de servidor (SLB), realizando apenas as tarefas de switching básicas, sem aceleração. Ambas as instâncias foram hospedadas no mesmo local, com configurações e hardwares idênticos. Com tudo acertado, o agente Gomez foi executado por cerca de 2000 interações em alguns dias, para obter uma amostra representativa.

Recursos e Benefícios da Tecnologia de Otimização BIG-IP

Os testes e os resultados descritos neste documento se aplicam aos produtos BIG-IP Application Accelerator e ao dispositivo BIG-IP Local Traffic Management, com os recursos apropriados de aceleração e otimização ativados. A seguir temos uma lista desses recursos e os benefícios que eles oferecem.

Fast Cache: Um cache de RAM interno para Web

- Alivia a carga de requisições dos servidores, ao servir páginas Web e objetos/imagens comuns no lugar dele.
- Acelera ainda mais o download de páginas quando combinado com a Intelligent Compression, ao permitir que objetos comprimidos entrem em cache e sejam servidos sem a latência de comprimir repetidamente o mesmo objeto.
- Armazena conteúdo comprimido e descomprimido ao mesmo tempo, e serve de forma inteligente o conteúdo correto, com base no que o cliente vai aceitar.
- Totalmente em conformidade com a RFC2616.
- Coloca em cache uma variedade de códigos de resposta de servidor: 200, 203, 206, 300, 301 ou 410.
- Pode responder a requisições condicionais GET e HEAD em lugar do servidor.
- Permite a configuração de vários repositórios de cache dedicados em um único sistema ("Multi-Store") para direcionar recursos de cache a aplicativos prioritários em um sistema compartilhado.
- Suporta a iRules, a linguagem de programação avançada da F5, para controle e gerenciamento superior sobre o conteúdo de cache.

Intelligent Compression: Transferência de carga de compressão sem agente

- Reduz o tempo de download de páginas – menos pacotes, tempo de envio e resposta menor. Reduz o consumo de largura de banda - serve ao mesmo número de usuários com menos largura de banda.
- Capacidade para direcionar compressão somente a clientes de conexão discadas ou de alta latência (longa distância) (Patente Pendente).

- Suporte nativo a todo navegador Web criado nos últimos cinco anos – nenhum plug-in ou software adicional de qualquer tipo é necessário.
- Transfere a carga de ciclos dos servidores e centraliza o gerenciamento para o processamento de compressão, oferecendo uma solução de menor custo, mais segura (por meio do direcionamento granular a clientes) e gerenciável.

TCP Express: Otimizações de TCP de última geração, líderes de mercado

- Numerosas técnicas de otimização de WAN e LAN que aceleram a transmissão de dados de acordo com as condições dos vários clientes e da rede.
- Centenas de melhorias de interoperabilidade de TCP entre pilhas disponíveis comercialmente (Windows 98, XP, 2000, IBM AIX, Sun Solaris e mais).
- A Otimização de WAN Centralizada / Sem Clientes inclui otimizações simétricas e assimétricas, sem downloads de clientes nem dispositivos adicionais.
- Baseados em otimizações de TCP de padrões abertos:
 - Delayed e Selective Acknowledgments (RFC 2018): Aumenta o desempenho ao lidar com pacotes perdidos e reordenados em WANs.
 - Explicit Congestion Notification (RFC 3168): Permite que o sistema BIG-IP sinalize os pontos proativamente, mostrando que os roteadores intermediários estão ficando sobrecarregados, para que eles possam voltar atrás e evitar a perda de pacotes.
 - Limited e Fast Re-Transmits (RFC 3042 e RFC 2582): Permite a retransmissão eficiente de dados perdidos para eliminar os efeitos de limite de tempo causados pela perda de pacotes.
 - Adaptive Initial Congestion Windows (RFC 3390): Estudos mostram um ganho de 30% para transferências de HTTP via satélite, e melhoria de 10% em conexões discadas de 28,8 bps, sem aumento na taxa de queda.
 - Slow Start with Congestion Avoidance (RFC 2581).
 - TCP Slow Start (RFC 3390): Permite uma maior utilização de largura de banda pelos links, para maior taxa de transferência em conexões públicas e linhas dedicadas de Internet existentes.
 - Controle de Atraso de Banda: Cálculo melhorado e expandido de atraso de largura de banda estima a carga ótima a ser exigida da rede sem exceder sua capacidade.
 - TimeStamps e Windows Scaling (RFC 1323): O BIG-IP permite o uso seletivo de marcas de tempo que acrescentam dados ao segmento TCP para auxiliar outras otimizações.

L7 Rate Shaping: Capacidade para priorizar uso de largura de banda

- Pode classificar (selecionar) o tráfego com informações L2 – L7 usando a iRules – isso quer dizer que o tráfego pode ser acompanhado com base no endereço MAC, informações de IP ou informações L7, como cookies HTTP.
- Contém classificações hierárquicas que permitem o empréstimo de banda entre categorias pais e filhas – por exemplo, vários clientes de FTP podem ter políticas de classificação únicas, mas se houver recursos disponíveis, pode ser desejável permitir o empréstimo.
- Pode classificar o tráfego de entrada e de saída independentemente.
- Suporta as disciplinas de fila Priority FIFO ("chegou primeiro, sai primeiro") e Stochastic Fair Queue ("aleatória").
- Taxas configuráveis de estouro, empréstimo e limite.
- Pode ser aplicado dinamicamente a servidores virtuais, na iRules.

**SSL Offload: Decodifica o SSL para que seus servidores não tenham que fazer isso**

- Aumenta a capacidade dos servidores - aliviando a carga do intenso processamento de SSL dos seus servidores, eles terão mais recursos para cuidar de seus objetivos específicos.
- Tempos menores de download de páginas – a F5 conta com os mais novos ASICs de criptografia de alto desempenho, capazes de lidar com dados criptografados de forma muito mais rápida que as CPUs comuns dos servidores.
- O dispositivo BIG-IP Application Accelerator suporta até 5.000 transações por segundo (TPS), e a poderosa plataforma BIG-IP LTM 6800 pode executar até 20.000 TPS. Oferece suporte único a novas conexões e criptografia em massa de dados, dentro de mecanismos especializados de transferência de carga por hardware.
- Pode reescrever de forma inteligente redirecionamentos de HTTP para HTTPS, para ajudar a integrar continuamente o SSL em seus aplicativos HTTP.
- A forte integração com a iRules permite decisões flexíveis de política, com base na força da criptografia, certificados dos clientes e quaisquer outras informações de SSL.
- Permite a inspeção profunda de tráfego de aplicativos e a modificação de tráfego já criptografado.

OneConnect™ TCP Offload: Reduz a carga de TCP multiplexando requisições HTTP

- Agrega várias conexões de cliente em menos conexões de servidor.
- Não atrasa requisições, nem as deixa em fila – mantém conexões de servidor suficientemente abertas para lidar com todas as conexões simultaneamente.
- Transforma cabeçalhos HTTP para incentivar conexões de servidor longas.
- Lida com conexões de cliente e de servidor de forma independente – isso permite uma incrível eficiência que pode transformar milhões de conexões em apenas algumas centenas para que sejam tratadas pelos seus servidores de back-end.

Content Spooling – Buffer de Servidor

- Capaz de ler respostas do servidor tão rapidamente quanto eles podem transmitir, eliminando o trabalho de se comunicar diretamente com clientes lentos.
- Transfere a carga do processamento de retransmissões e otimiza fluxos individuais, para obter o melhor desempenho para cada usuário final, enviando dados tão rapidamente quanto eles conseguem receber.
- Com dados lidos e tirados do servidor mais rapidamente, o servidor fica livre para processar mais conexões, o que aumenta sua capacidade.