

O mito da banda e o desempenho dos aplicativos

Visão Geral A lei de Moore determina que a densidade dos dados dobra aproximadamente a cada 18 meses, e a lei de Metcalfe diz que o valor de uma rede aumenta ao quadrado do número de usuários. Como esses postulados se comprovam na prática, as companhias globais descobriram a vantagem de integrar a tecnologia da informação em cada aspecto de suas operações, e o setor de serviços de comunicação mundial de dados hoje gera uma receita de mais de 19 bilhões de dólares todos os anos, dos quais uma parte crescente é derivada dos serviços VPN IP.

Apesar do crescimento mundial da demanda por largura de banda, o suprimento ultrapassou a demanda por uma ampla margem. Durante a rápida expansão da Internet na década de 90, o setor de comunicação de dados criou uma infra-estrutura capaz de distribuir banda barata em altos volumes. De fato, a banda se tornou tão abundante que mesmo os efeitos da lei de Metcalfe ainda são insuficientes para consumir a capacidade disponível por muitos anos. O resultado desse desequilíbrio foi a massificação da banda, o rápido declínio do preço e um ambiente de fornecedores que promoviam ativamente o mito de que grandes quantidades de banda podem resolver qualquer problema de desempenho.

Mas, conforme as implementações de aplicativos corporativos foram expandidas para a área de longa distância, um ambiente em que a banda às vezes é tão abundante quanto nas redes locais, os gerentes de TI testemunharam um declínio radical no desempenho de aplicativos. Eles se perguntam: "Por que duas redes, a LAN e a WAN, com capacidades de banda idênticas, oferecem resultados com desempenho tão diferente?"

A resposta é que o desempenho dos aplicativos é afetado por muitos fatores associados tanto com a rede quanto com a lógica do aplicativo, e que isso precisa ser resolvido para que se obtenham resultados satisfatórios em termos de desempenho de aplicativos. No nível da rede, o desempenho dos aplicativos é limitado pela alta latência (efeito da distância física), jitter, perda de pacotes e congestionamento. No nível do aplicativo, o desempenho é ainda mais limitado pelo comportamento natural dos protocolos de aplicativos (especialmente em situações com latência, jitter, perda de pacotes e congestionamento no nível de rede), que executam handshakes em excesso nos links da rede, e pela serialização dos próprios aplicativos.

Este documento tenciona esclarecer os problemas que afetam o desempenho dos aplicativos na área de longa distância e oferecer aos gerentes de TI o conhecimento necessário para planejar soluções estratégicas de implementação e aceleração de aplicações corporativas.

Mitos comuns do desempenho de aplicativos

Mito nº 1: O desempenho dos aplicativos depende apenas da largura de banda
O desempenho e a capacidade dos aplicativos são influenciados por muitos fatores. A latência e a perda de pacotes têm um efeito profundo no desempenho de aplicativos. A Lei de Little, uma descrição seminal da teoria de enfileiramento e uma equação que modela os efeitos da distância física (latência) e perda de pacotes, ilustra o impacto desses dois fatores no desempenho dos aplicativos.

Essa lei determina que:

$$\text{Lambda (capacidade)} = n \text{ (número de solicitações em aberto)} / t \text{ (tempo de resposta)}$$

Em termos de protocolos baseados em IP, isso significa:

$$\text{Capacidade do TCP} = \text{tamanho da janela de congestionamento} / \text{tempo de ida e volta (rtt)}$$

Portanto, conforme aumenta o tempo de ida e volta (RTT) de cada solicitação, a janela de congestionamento deve aumentar ou a capacidade do TCP irá diminuir. Infelizmente, o TCP não gerencia janelas grandes de forma eficiente. Como resultado, mesmo pequenas quantidades de latência e perda de pacotes podem derrubar o desempenho de rede para um determinado aplicativo para menos de 1 megabit por segundo. Mesmo se a capacidade de banda fosse aumentada para 100 Mbps, o aplicativo jamais consumiria mais do que 1% da capacidade total. Nessas condições, os administradores que aumentam a capacidade da rede desperdiçam dinheiro em um recurso que não pode ser consumido.

"O comportamento macroscópico do algoritmo Congestion Avoidance do TCP", por Mathis, Semke, Mahdavi e Ott, publicado no Computer Communication Review nº 27(3) em julho de 1997, oferece uma fórmula útil e simples para o limite superior da taxa de transferência:

$$\text{Taxa} = (\text{MSS}/\text{RTT}) * (1 / \sqrt{p})$$

Onde:

Taxa: a taxa de transferência ou capacidade do TCP

MSS: o tamanho máximo do segmento (fixo para cada rota da Internet, normalmente, 1460 bytes)

RTT: o tempo de ida e volta (medido pelo TCP)

p: a taxa de perda de pacotes

A figura abaixo ilustra esse conceito:

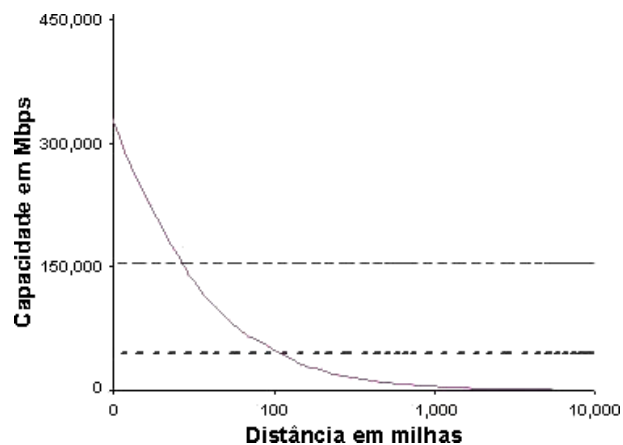


Figura 1: o desempenho do TCP em relação à distância física

Nas redes de longa distância (WANs), as fontes de tempos altos (latência) de ida e volta incluem distância física, padrões ineficientes de roteamento de rede e congestionamentos na rede, elementos abundantes nas WANs.

Hoje, muitas pilhas do protocolo TCP são ineficientes no que diz respeito ao gerenciamento das retransmissões. De fato, algumas implementações podem ter de

retransmitir toda a janela de congestionamento se um único pacote for perdido. Elas também tendem a recuar exponencialmente (ou seja, reduzir a janela de congestionamento e aumentar os temporizadores de retransmissão) quando ocorre congestionamento na rede, um comportamento que é detectado pelo TCP como perda de pacotes. E, embora a perda normalmente seja insignificante em redes de frame relay (menos de 0,01% em média), ela é muito significativa nas redes VPN IP que vão e voltam a mercados como a China, onde as taxas de perdas normalmente superam os 5%. Nesse último cenário, as altas taxas de perda podem ter um efeito catastrófico no desempenho.

Quando a perda de pacotes e os efeitos da latência são combinados, a queda de desempenho é ainda mais severa. A figura abaixo ilustra esse conceito:

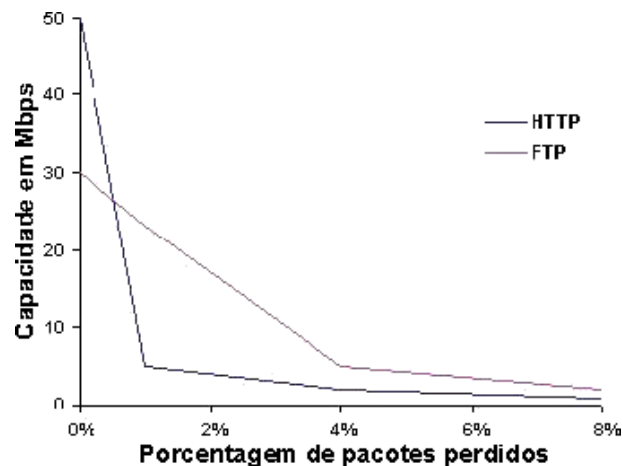


Figura 2: desempenho do TCP na presença de perda de pacotes

Mito nº 2: O TCP exige recuos agressivos para garantir o equilíbrio

Muitos engenheiros de rede acreditam que um recuo agressivo frente ao congestionamento é necessário para manter equilibrado o acesso à rede. Embora em alguns casos isso seja verdade, em outros, não é. Quando o controle de congestionamento é responsabilidade de cada host em uma rede, ambiente em que cada host não tem conhecimento das necessidades de banda dos outros hosts, os recuos agressivos podem ser necessários para garantir o equilíbrio da rede. Entretanto, se o congestionamento é gerenciado na rede, por um sistema que vê todo o tráfego em uma determinada conexão WAN, é possível obter uma capacidade muito maior e mais eficiente, e os recuos agressivos não são necessários.

O comportamento-padrão do protocolo especifica que, quando um host consome banda, ele deve fazer isso independentemente de:

- Requisitos do aplicativo
- Quantidade de banda disponível
- Volume de competição que existe por essa banda

O resultado é uma situação em que os aplicativos sofrem escassez de recursos de banda ao mesmo tempo em que a rede é, em sua maior parte, subutilizada. Obviamente, essa situação é muito ineficiente.

Uma solução muito melhor para o problema do equilíbrio do TCP é permitir que os hosts individuais consumam tanta banda quando precisarem, desde que todos os outros hosts recebam os serviços adequados quando precisarem deles. Isso pode ser feito com a implementação de uma janela de congestionamento única,

compartilhada por todos os hosts e gerenciada na própria rede. O resultado é um sistema em que os hosts obtêm a banda de que necessitam em períodos de pouca competição, e todos os hosts recebem banda suficiente quando a competição é mais intensa.

Esse método de janela única oferece uma utilização mais alta e uma capacidade geral maior, de maneira consistente. Cada host vê uma rede limpa, rápida e que nunca perde pacotes (e que, portanto, não diminui o desempenho do TCP. Veja o mito nº1), e as demandas cumulativas de tráfego são combinadas à capacidade de buffering geral da rede. Como resultado, os gerentes de TI têm uma utilização otimizada das redes, sob as mais variadas condições de latência e perda.

As soluções de janela única podem ser criadas de forma completamente transparente aos sistemas clientes. Os componentes de tais soluções podem incluir tecnologias TCP, como ACKs seletivos, gerenciamento da janela de congestionamento local, algoritmos de retransmissão melhorados e dispersão de pacotes. Essas capacidades são então combinadas com outras tecnologias que combinam as exigências de capacidade dos aplicativos à disponibilidade dos recursos de rede e que rastreiam os requisitos de banda de todos os hosts na rede. Ao agregar a capacidade de vários links WAN paralelos, essa tecnologia pode oferecer confiabilidade e capacidade ainda maiores.

Mito nº 3: A compressão de pacotes melhora o desempenho dos aplicativos

Embora as técnicas comuns de compressão de pacotes possam reduzir a quantidade de tráfego na WAN, elas geralmente diminuem o desempenho dos aplicativos, já que adicionam latência às transações. Essas técnicas exigem que os pacotes sejam enfileirados, compactados, transmitidos, descompactados no receptor e então retransmitidos, e isso pode consumir muitos recursos e gerar muita latência, tornando mais lentos os aplicativos que precisam de aceleração.

As soluções de desempenho de aplicativos da próxima geração combinam a simplificação do protocolo com técnicas transparentes de redução de dados. Comparadas às soluções baseadas em pacotes, as soluções da próxima geração diminuem em muito a quantidade de dados que precisam ser retransmitidos, eliminando a latência que é introduzida pelo comportamento do protocolo por causa da distância física, podendo levar o desempenho das redes de longa distância a velocidades de gigabits. As técnicas transparentes de redução de dados geralmente incluem múltiplos dicionários, em que o dicionário de nível 1 é pequeno e altamente eficiente na redução de pequenos padrões de dados, e o dicionário de nível 2 é um espaço de vários gigabytes que pode ser usado para reduzir padrões muito maiores.

Mito nº 4: A tecnologia de qualidade do serviço acelera os aplicativos

A qualidade do serviço (QoS), se usada corretamente, é uma tecnologia muito benéfica que pode ajudar a gerenciar o desempenho de aplicativos. Entretanto, a única coisa que a QoS pode fazer é dividir a banda existente em múltiplos canais virtuais. A QoS nada faz para mover mais dados ou simplificar o comportamento do protocolo. Ela simplesmente decide, de maneira inteligente, quais pacotes serão descartados. E, embora seja melhor descartar pacotes de maneira controlada do que deixar isso ao acaso, descartar pacotes não acelera os aplicativos.

Muitas implementações QoS dependem de números de porta para rastrear aplicativos. Como os aplicativos podem negociar a definição de portas de forma dinâmica, esses mecanismos devem ser configurados para reservar grandes faixas de portas para garantir a cobertura das portas de fato utilizadas pelo aplicativo.

Para que a QoS seja mais eficiente, ela deve ser dinâmica. As implementações QoS de primeira geração dividiam links grandes em múltiplos links menores, reservando banda de maneira estática, havendo ou não necessidade. A canalização de uma rede, feita dessa forma, pode garantir a disponibilidade de banda para aplicativos críticos como voz, mas, na realidade, desperdiça banda, porque ela está reservada para aplicativos específicos, mesmo quando esses aplicativos não estão em uso.

As soluções QoS dinâmicas, por outro lado, garantem que a banda seja reservada apenas quando os aplicativos podem usá-la. Um uso comum dessa tecnologia é ampliar as janelas de back-up das empresas, habilitando o back-up contínuo de dados quando houver banda disponível.

Solução

A F5 reúne tudo isso

As soluções de aceleração de aplicativos da F5 oferecem um aumento dramático no desempenho de aplicativos e reduzem muito os custos da WAN. A F5 oferece esses benefícios monitorando os efeitos limitadores nas condições da rede, ajustando o comportamento do protocolo e gerenciando todos os níveis da pilha de protocolos, desde a camada de rede até a camada de aplicação.

Mais especificamente, a F5 integra tecnologias avançadas de aceleração de transporte, como a aceleração adaptativa TCP, a redução transparente de dados e a qualidade de serviços perceptiva a sessões, com as melhores tecnologias de aceleração de aplicativos, incluindo caching dinâmico de objetos, proxies inteligentes de aplicativos e criptografia inteligente de aplicativos. O sistema tem o suporte das funções de geração de estatísticas e monitoramento, que permitem o gerenciamento em tempo real do comportamento da rede e de aplicativos.

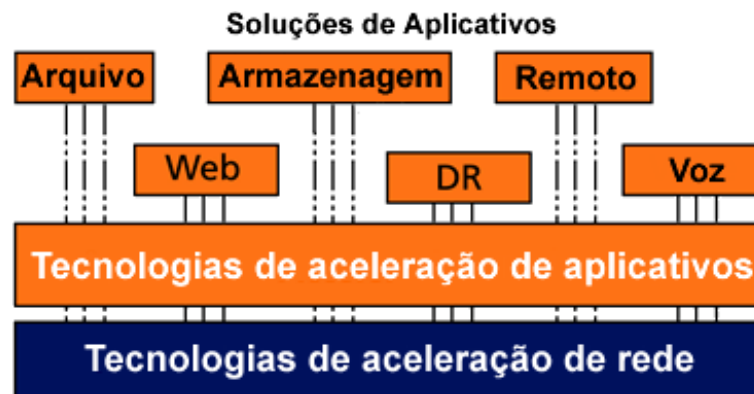


Figura 3: Arquitetura

A F5 oferece um desempenho de aplicativos WAN similar àqueles da LAN. As soluções da F5 aceleram aplicativos como ERP, CRM, e-mail, transferência de arquivos, replicação de dados e outras aplicações, resultando em um desempenho rápido e previsível para todos os usuários da WAN.

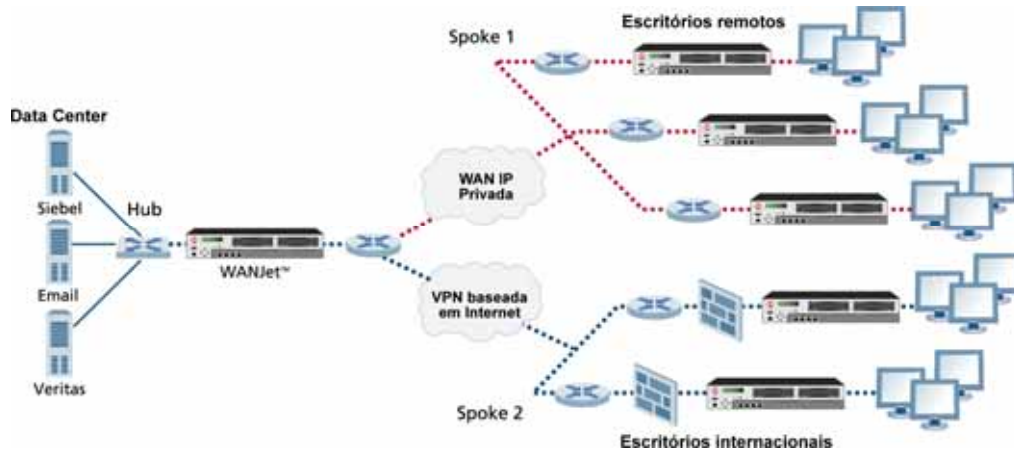


Figura 4: Implementações em duas pontas aceleram todo o tráfego de aplicativos na WAN

As soluções de aceleração e otimização da WAN da F5 são implementadas em dispositivos de hardware da F5. O modelo de data center da F5, o WANJet 500, traz tolerância a falhas, escalabilidade maciça e desempenho de até 622 Mbps. Para implementação em filiais, o WANJet 200 traz tolerância a falhas, operação silenciosa e desempenho de até 2 Mbps.

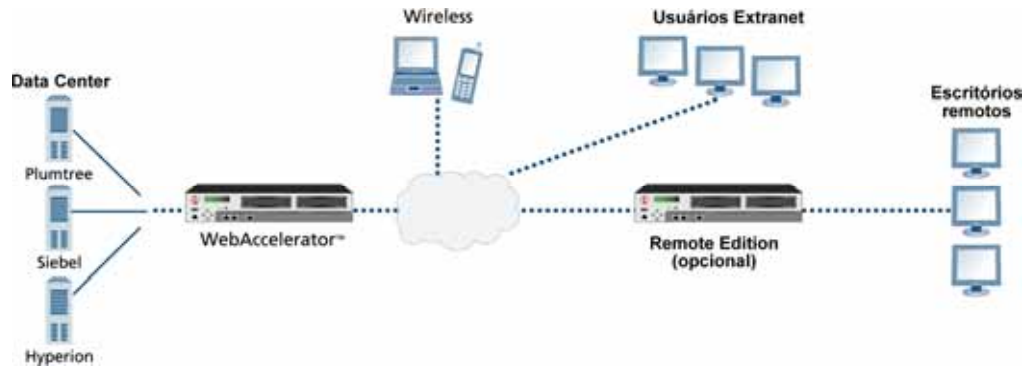


Figura 5: implementações de ponto único aceleram o desempenho de aplicativos web corporativos como SAP, Siebel, Oracle e portais corporativos

Resultados típicos de desempenho

Desempenho do TCP sem aceleração

Neste exemplo, um cliente Windows XP está usando FTP ativo para obter um arquivo de 10 MB de um servidor FTP Linux Redhat 7.3. O link é uma linha E1 de 2 Mbps, com 400 milissegundos de tempo de ida e volta, com as condições aproximadas de um link da Califórnia para a Ásia.

FTP Get	Taxa de Transferência	Tempo de Transferência	Utilizaçã o do Link
0% de perda de pacotes	50 KB/s	250 segundos	20%
1% de perda de pacotes	20 KB/s	625 segundos	8%

Neste cenário, é claro que uma única transferência FTP não pode alcançar mais do que 20% de utilização. Com 1% de perda de pacotes, a perda de desempenho é de mais de 50%. Adicionar banda a esse link não aumentará a capacidade de forma alguma.

Desempenho do TCP com WANJet, TDR desabilitado

Transferindo o mesmo arquivo pelo dispositivo WANJet, sobre uma rede com 0% de perda de pacotes, temos como resultado um aumento de desempenho – cinco vezes maior do que o desempenho nativo. Em uma rede com 1% de perda, o desempenho é doze vezes melhor.

FTP Get	Taxa de Transferência	Tempo de Transferência	Utilização do Link
0% de perda de pacotes	230 KB/s	45 segundos	100%
1% de perda de pacotes	200 KB/s	50 segundos	100%

Desempenho do TCP com WANJet, TDR habilitado

Transferindo novamente o mesmo arquivo pelo dispositivo WANJet, mas agora com os algoritmos patenteados de gerenciamento de congestionamento e redução transparente de dados da F5, temos um aumento de desempenho que varia entre 40 e 625 vezes. O primeiro exemplo mostra uma transferência de um arquivo de 10 MB compactável a um terço do tamanho, melhorando o desempenho em 40 vezes. O segundo exemplo mostra a transferência do mesmo arquivo, dessa vez com metade dos bytes modificados, aumentando o desempenho em 78 vezes. O terceiro exemplo mostra uma leitura subsequente do mesmo arquivo, com um ganho de desempenho da ordem de 625 vezes.

	Taxa de Transferência	Tempo de Transferência	Utilização do Link	Aumento do Desempenho
Arquivo de 10 MB, compactável a 1/3 – FTP get ativo	700 KB/s	15 segundos	100%	40x
Arquivo de 10 MB, compactável com dados modificados	1.500 KB/s	6 segundos	100%	100x
Arquivo de 10 MB, compactável a 1/3, com dados modificados	10.300 KB/s	1 segundo	1%	625x

Note que essas transferências mostram taxas de dados efetivas que são de 3 a 44 vezes maiores do que a taxa efetiva de uma linha E1 nativa.

Conclusão

O desempenho dos aplicativos na WAN é afetado por um grande número de fatores além da banda. A noção de que a banda resolve todos ou mesmo a maioria dos problemas de desempenho dos aplicativos é um mito. No nível da rede, o desempenho dos aplicativos é limitado pela alta latência, jitter, perda de pacotes e congestionamento. No nível de aplicação, o desempenho é limitado por fatores como: o comportamento natural dos protocolos do aplicativo, que não foram criados para condições de WAN; protocolos de aplicativos que executam handshakes excessivos e a serialização dos próprios aplicativos.

As soluções de aceleração de aplicativos da F5 reconhecem a interdependência crítica entre o comportamento do nível dos aplicativos e do nível de transporte. As soluções da F5 oferecem um desempenho previsível de aplicativos, a capacidade ampliada de 3 a mais de 500 vezes e desempenho aprimorado de aplicativos em redes diferentes, desde aquelas com gerenciamento de qualidade superior até as redes VPN IP de massa. As vantagens arquitetônicas da F5 resultam em soluções de desempenho de aplicativos que oferecem o melhor desempenho em escalabilidade maciça e um retorno de investimento medido em meses.



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 consumam 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).