Realtime publishers

The Definitive Guide to Cloud Acceleration

Dan Sullivan

sponsored by

CDNetworks

Chapter 6: How to Choose a Cloud Application Acceleration Vendor 80
Global Reach 81
Technical Dimension of Global Reach 81
Business Dimension of Global Reach 82
Dynamic Content Acceleration 83
High Availability 84
Faster Application Performance 85
Better End User Experience 85
Security Considerations 85
SSL Encryption and Cloud Acceleration 86
Need for Encryption vs. Cost 86
Accelerating Encryption 86
Distributed Denial of Service (DDoS) Protection 87
The Structure and Function of a DDoS Attack 87
Targets of DDoS Attacks 89
Responding to DDoS Attacks 90
Data Security 91
Authentication 91
Architecture Considerations 91
Key Business Considerations 92
Impact of Slow Applications 93
Adverse Customer Experiences 93
Summary 94

Copyright Statement

Chapter 6: How to Choose a Cloud Application Acceleration Vendor

Throughout, this guide has examined the challenges to delivering services from the cloud, with particular attention to the design of Web applications and the architecture of the Internet. Providing high-performance applications to a geographically distributed user base is a challenge for several reasons, including the physics and engineering constraints of networking as well as the logical design of Internet protocols. Previous chapters describe techniques to increase throughput and reduce latency using a combination of distributed content servers and higher-performance protocols. This chapter turns the focus to evaluation criteria for selecting a cloud application acceleration provider.

Cloud application acceleration can improve the performance of your applications but only if essential components and functionality are in place. In addition to strictly performance-related criteria, it is important to consider factors such as content life cycle management and security. This chapter is organized into several sections, each of which addresses a key evaluation area:

- Global reach
- Dynamic content acceleration
- Security
- Architecture considerations
- Key performance metrics
- Technical support
- Key business considerations

These evaluation areas are generally applicable to enterprise applications, but some topics might be more important than others. Application and organization requirements should determine the weight applied to each of these areas. For example, if your primary concern is increasing the performance of analytic applications used by customers across the globe, dynamic content acceleration is more important than is static content caching. Consider the suite of applications your organization is supporting as you determine the relative importance of each of these areas. Also keep in mind strategic plans and their implications for system design. You might find that you have or will likely have a combination of applications that could benefit from cloud application accelerations services.

Global Reach

Global reach in the context of cloud application acceleration has both a technical and a business dimension. Consider both during your evaluation.

Technical Dimension of Global Reach

Companies turn to cloud application acceleration and content distribution network providers to improve the user experience for application users. Ideally, all users would have a high-quality experience using a responsive system with high availability, low latency, and accurate, up-to-date content. Users in North America accessing content that originates in Europe should have the same approximate experience as Europeans accessing the same content. Content delivery networks and network acceleration on a global scale are both required to meet this objective.

Content delivery networks are geographically distributed servers that maintain local copies of content that originates elsewhere. For example, a server in the eastern United States might host a content management system (CMS) with static content used in an organization's web site. Customers from the eastern US and Midwest will likely receive this content with low latency. Users in the western parts of the country might experience somewhat higher latency than their counterparts in the East. Users in Asia, however, would have to wait significantly longer for a static Web page because transferring content across North America and the Pacific Ocean will take significant time.

As you evaluate your need for global reach, consider your existing and potential user base. Will your organization be implementing new strategies and services in Europe and Asia? If so, how can a content delivery network support those services? What is the average latency from the edge servers that cache content from originating servers to different parts of your market (see Figure 6.1)?

Figure 6.1: Content delivery networks require globally distributed edge servers to support delivery from caches closer to end users.

Business Dimension of Global Reach

In addition to the technical aspects of caching content in edge servers around the globe, there are legal and cultural considerations to keep in mind. Governing bodies around the world have varying restrictions on Internet content and their own requirements with regards to registering and licensing Internet service providers (ISPs). These regulations might be minimal and easily addressed in consultation with local legal counsel. At the other end of the spectrum, you might find frequent and ongoing government regulation and monitoring. China, for example, has established relatively strict controls on the Internet within the country (see Figure 6.2).

Figure 6.2: Popular sites outside of China and those posting content deemed inappropriate might be blocked by the Chinese government. The censoring infrastructure is commonly known as the Great Firewall of China.

Consider how content delivery network and network acceleration providers can assist you with navigating local regulations, complying with operational restrictions and responding to orders from the government. As with many other business services, organizations might consider developing in-house expertise to manage these issues. This choice is reasonable in some cases—for example, when you have a long history of business in the country, have in-depth knowledge of legal and cultural issues, and understand legal procedures in the country. When the cost of developing expertise in local matters outweighs the benefits, it is appropriate to consider how your content delivery network provider and network acceleration provider might be able to assist you with local matters.

Global reach encompasses both technical aspects, such as the distribution of edge servers and the ability to accelerate network traffic over global distances, and business aspects, such as support for complying with local regulations around the globe. Let's next turn to a more in-depth look at several technical areas.

Dynamic Content Acceleration

There are many types of applications that lend themselves to protocol optimizations instead of caching. These include order-processing applications, such as airline ticket reservation systems, and ad hoc query applications, such as business intelligence systems.

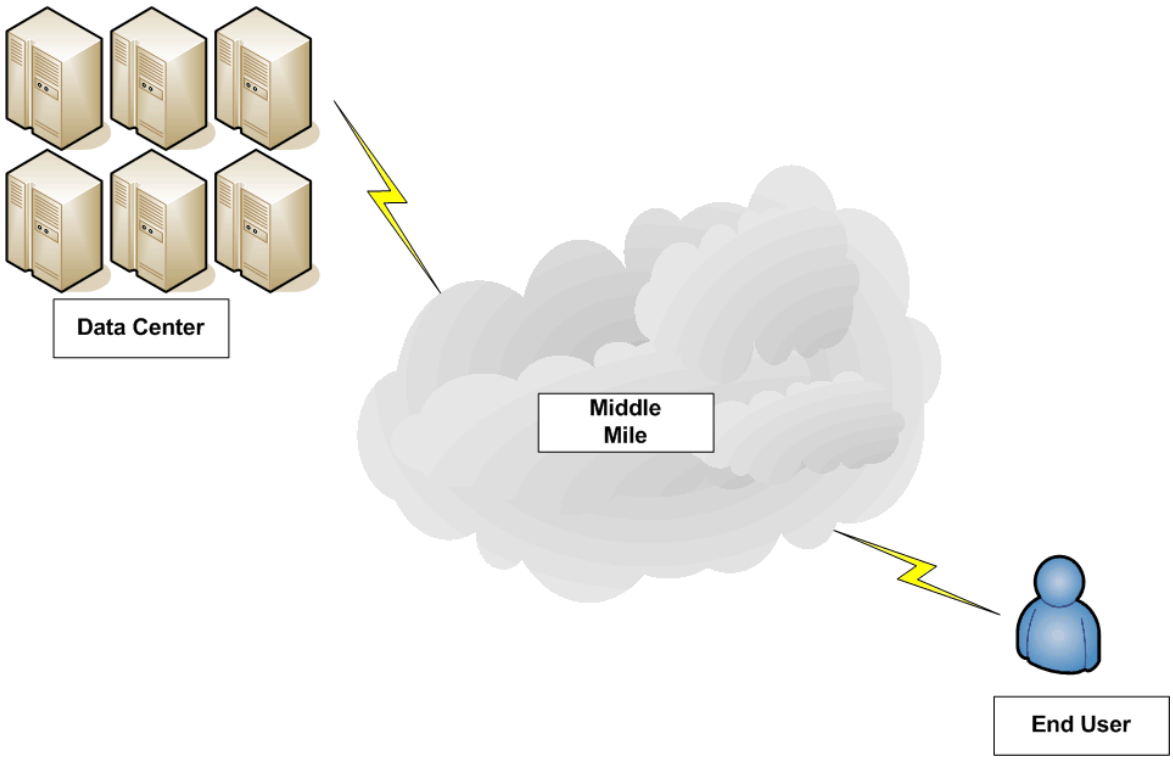

Figure 6.3: The middle mile is the intermediate network infrastructure between edge networks.

When evaluating dynamic content acceleration techniques, consider how your provider implements acceleration. For example, traffic in the middle mile of the network between two edge servers can be optimized because the acceleration network provider controls both endpoints. The servers can negotiate protocol settings that reduce the number of packets that must be resent when packets are dropped and set other TCP configuration parameters to reduce overhead on the network.

The distribution of edge servers is also a key consideration in evaluating dynamic content acceleration. Edge servers should be distributed in ways that reach large portions of the user base, mitigate the impact of poor performance peering agreements between ISPs, and maintain reasonable loads on the servers.

The implementation choices made by cloud acceleration providers are important because they can impact key application requirements, including:

- High availability
- Faster application performance
- Improved end user experience

High Availability

Customers expect business Web sites and applications to be available 24x7 in spite of hardware failures, network problems, and malicious cyber-attacks. Cloud acceleration providers can help mitigate the risk of unwanted downtime by providing high-availability hardware and networks.

Hardware fails. Today's large-scale data centers house large numbers of servers and storage devices and therefore it is reasonable to assume that at least one component in a large data center will fail in production. Cloud acceleration vendors can provide for high-availability Web sites and applications with a combination of failover clusters, redundant storage, and multiple data centers.

When high availability is a requirement for an application, that application may be run in a cluster of servers. In some cases, all servers in a cluster share the application workload and if one fails, the other servers will continue to process the workload. In other cases, a single server may process the full workload while a stand-by server is constantly updated with the state of the primary server. If the primary server fails, the stand-by server takes over processing the workload. Cloud providers should support the appropriate type of high-availability configuration appropriate for your applications.

Storage systems are also subject to occasional failures. Redundant storage systems improve high availability by reducing the chances that a storage failure will lead to application downtime.

Large-scale network problems and natural disasters can result in large-scale disruptions to a data center. Cloud acceleration providers with multiple data centers can continue to provide access to applications by routing traffic away from the disrupted data centers to other data centers hosting the same content or able to run the applications that had been available from the disrupted data center.

Faster Application Performance

The responsiveness of a Web application is determined, in part, by network latency. If large numbers of packets must be exchanged between a server and a client, application performance can suffer. TCP and HTTP protocols can both require multiple round-trip packet exchanges between clients and servers. Dynamic acceleration technologies can reduce network overhead by optimizing the number of packets exchanged between client devices and servers.

Network acceleration is particularly important in the "middle mile," which can constitute the longest segment between a server and a client device. Reducing the amount of time required to exchange data on the longest segment can substantially reduce the time users are waiting for a response from a server. Also, network acceleration systems can maintain pools of TCP connections that can be reused without incurring the overhead of creating a new TCP connection.

Better End User Experience

The most important beneficiaries of Web application acceleration are end users. Users of ecommerce applications will find that dynamic acceleration improves responsiveness and allows users to complete transactions more efficiently. Analysts working with data warehouses and business intelligence applications can perform more analytic operations in less time when network traffic is accelerated.

Dynamic content acceleration allows businesses to provide highly available, responsive Web applications that deliver an improved end user experience.

Security Considerations

Security is a broad topic that encompasses confidentiality, integrity, and availability of data and applications and applies to virtually all aspect of information technology. Cloud acceleration is no exception. Several topics are of particular importance with regards to content distribution networks and dynamic content accelerations:

- Secure Sockets Layer (SSL) encryption and cloud acceleration
- Distributed Denial of Service (DDoS) protection
- Data security
- Authentication

SSL Encryption and Cloud Acceleration

SSL, predecessor of the newer Transport Layer Security (TLS), is the commonly used protocol for encrypting data transferred over the Internet. It is of particular importance in protecting the confidentiality of content as it is transmitted.

As you plan to deploy content delivery networks and dynamic content acceleration, consider which content should be encrypted during transmission. If you organization has a data classification system in place, that system can inform you about the types of content that might need to be encrypted. For example, data subject to government or industry regulations may require strong encryption anytime it is transmitted over the Internet. In other cases, data classified as public—that is, data that if released to the public would not cause any harm to the organization—can be transmitted without encryption.

Need for Encryption vs. Cost

One way to address the issue of deciding which data to encrypt is to simply encrypt all data. This approach sounds prudent at first glance. After all, "better safe than sorry" is one way to address security questions. The problem is that this approach does not take into account the cost of encrypting data.

SSL encryption uses a combination of two encryption techniques: asymmetric cryptography and symmetric cryptography. Both use keys to encrypt and decrypt data. Symmetric key cryptography uses one key while asymmetric cryptography uses two. Symmetric key cryptography is the less computationally demanding of the two methods but because only one key is used, partners exchanging encrypted data have to share a common key. Transmitting an encryption key in unencrypted form is risky and can lead to a compromised key. Asymmetric cryptography is computationally expensive but has the advantage of not requiring a shared key.

Accelerating Encryption

SSL takes a best-of-both-worlds approach and uses both asymmetric and symmetric key cryptography. Asymmetric cryptography is used during the SSL handshake (see Figure 6.8) when two devices are establishing an encrypted session. During the handshake, the two devices exchange information about the algorithms and other parameters that each supports. Asymmetric techniques are used, so this communication can occur over a secured channel. During the handshake, the devices exchange a symmetric key that is used to encrypt data for the rest of the session.

Encrypting data, especially during the handshake using asymmetric encryption, is computationally demanding. Encrypting large volumes of data over many different sessions can place heavy demands on CPUs. Edge servers providing access to encrypted data can benefit from SSL acceleration.

SSL acceleration is implemented by specialized hardware designed to offload computation from server CPUs. Content delivery network providers may or may not provide SSL encryption, so be sure to evaluate support for SSL acceleration if you plan to use SSL to protect the confidentiality of your data.

Distributed Denial of Service (DDoS) Protection

Businesses and other organizations face an array of security threats; one of the most challenging is the Distributed Denial of Service (DDoS) attack. As the name implies, the object of the attack is to disrupt services provided by legitimate Web sites and applications so that they are not available to customers and users. There are a few common types of DDoS attacks, but they all involve overwhelming servers with malicious network traffic.

The Structure and Function of a DDoS Attack

There are multiple ways to implement DoS attacks, including flooding servers with network traffic and disrupting DNS functions. A simple type of DDoS attack sends large volumes of SYN or other types of TCP packets that prompt the receiving server to open connections. When a legitimate connection is open, the client device that initiated the connection will respond after it receives an acknowledgement packet from the server (see Figure 6.5).

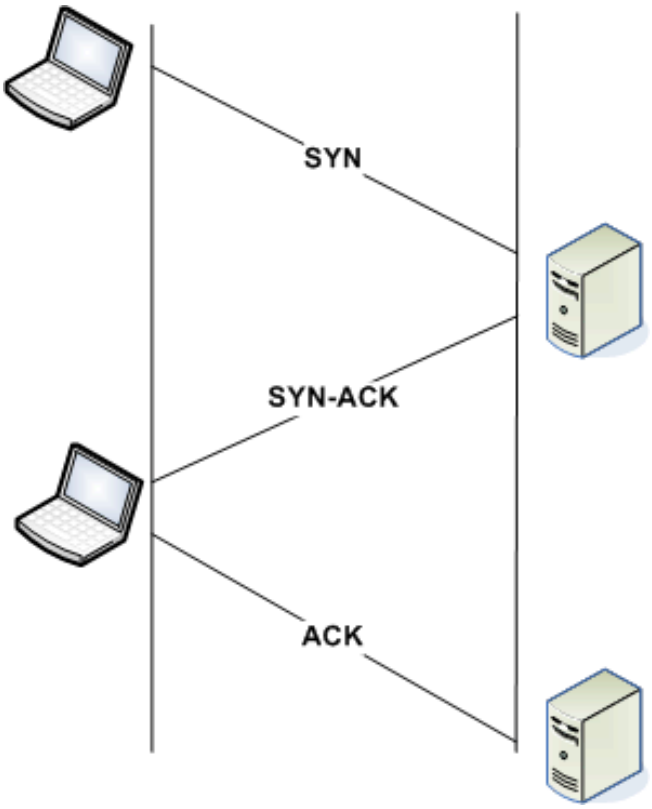

Figure 6.4: In a typical TCP handshake, a client initiates a connection with a SYN packet, the server responds with a SYN-ACK and the client then responds with an ACK.

In the case of a malicious attack, the client does not respond and the server is left holding the connection open while it waits for a response. Eventually, the connection will time out, but the during that time, the attacker will have issued other connection requests ultimately consuming all connection resources. As a result, legitimate traffic is unable to establish connections to the server (see Figure 6.5).

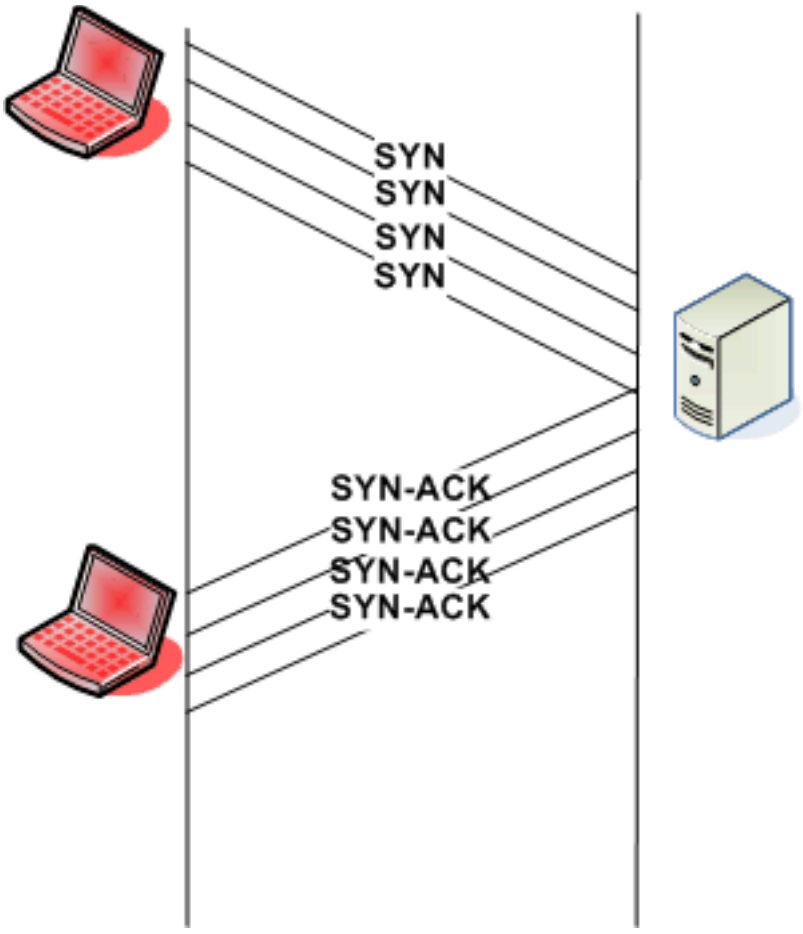

Figure 6.5: In a malicious DoS attack, the clients flood the server with SYN packets leading to corresponding SYN-ACK, which are unacknowledged by the attacker. As a result, the server waits for acknowledgement while connection resources are consumed for non-functional connections.

DDoS attacks use a collection of compromised devices, known as a botnet, to flood a target server (see Figure 6.6). The compromised devices have been infected with malware that allows the attacker to issue commands to the compromised computers. These commands specify the type of attack to launch and the target server. In addition to the compromised computers that are flooding servers with malicious traffic, botnets often include multiple command and control servers. The person controlling the botnet, known as the bot herder, communicates with command and control servers, which in turn communicate with compromised devices.

One way to disrupt a botnet is to shut down or isolate the command and control server so that it can no longer issue commands. Botnet designers have recognized this potential single point of failure and have developed techniques to support multiple command and control servers. If one is identified and taken offline, another can assume the responsibilities of communicating with compromised devices. As a result, botnets are resilient to attacks on their infrastructure.

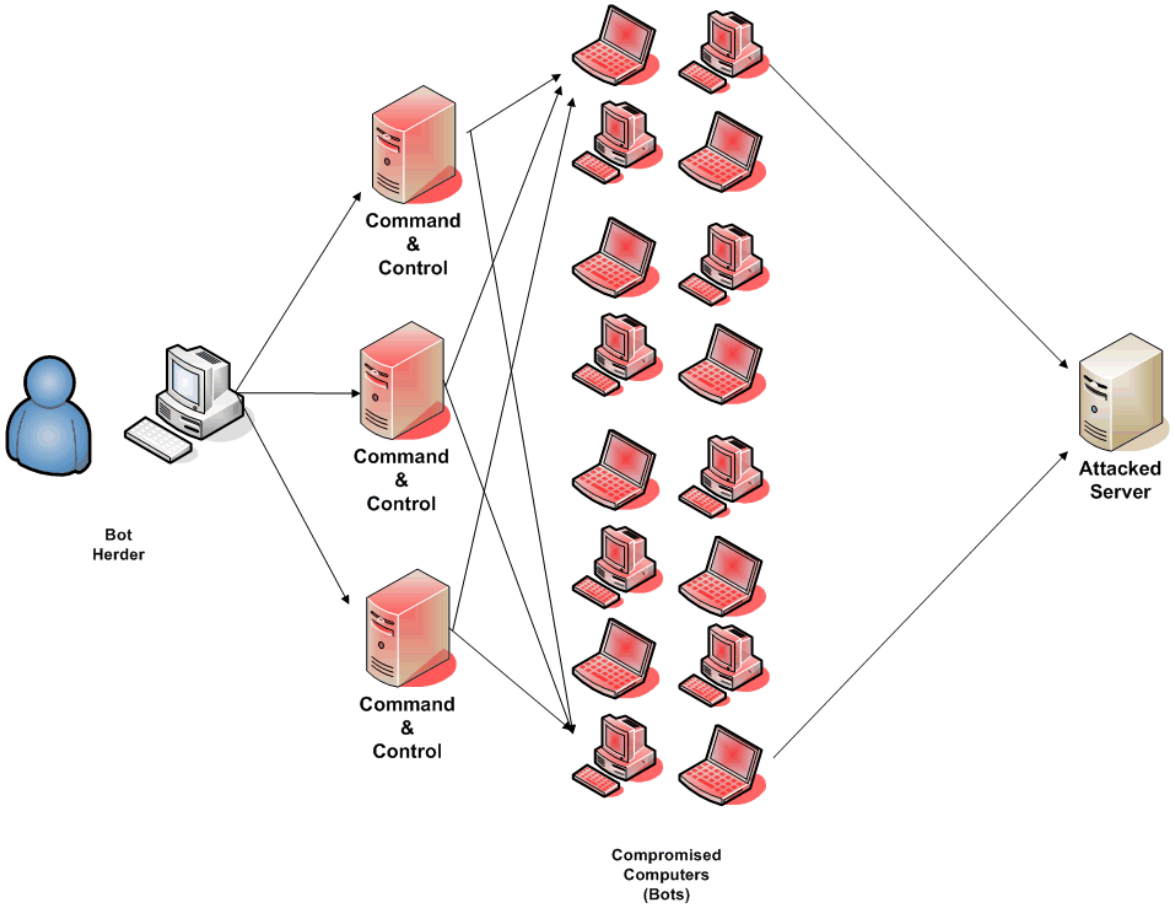

Figure 6.6: Botnets are distributed systems controlled by multiple command and control systems, making them difficult to disrupt by taking down bots or command and control servers.

Targets of DDoS Attacks

DDoS attacks are simple but effective. Government agencies, financial institutions, and retailers have all been victims of DDoS attacks. The impact on businesses is multifaceted. Retailers suffer immediate adverse effects, including lost sales. All organizations may suffer damage to their brand as users are unable to conduct normal business operations with the victims of DDoS attacks.

One of the reasons for the growing threat of DDoS attacks is that they are relatively easy to launch. Information on how to launch a DDoS attack is readily available online. DDoS application code is available as well. Even those without the technical skill to implement their own attack can find DDoS "service providers" on the cybercrime black market who have their own DDoS infrastructure and launch attacks for others.

Responding to DDoS Attacks

One method to respond to the threat of DDoS attacks is to add infrastructure to absorb an attack. This approach is not practical. Attackers can launch attacks consuming 40 to 60GB of bandwidth. They have access to multiple compromised devices, many of which have high-speed Internet connections, so it is a fairly simple matter to scale up the size of botnets to consume all available network and server resources at the target site.

A better method of responding to DDoS attacks is to use DDoS absorption techniques, as Figure 6.7 illustrates. DDoS absorption systems are network devices that analyze traffic and detect patterns indicative of a DDoS attack. Malicious packets are filtered before they reach production servers on a network. Depending on the type of attack, network engineers might use data from the DDoS absorption systems to determine whether some types of requests should be blocked or blacklisted.

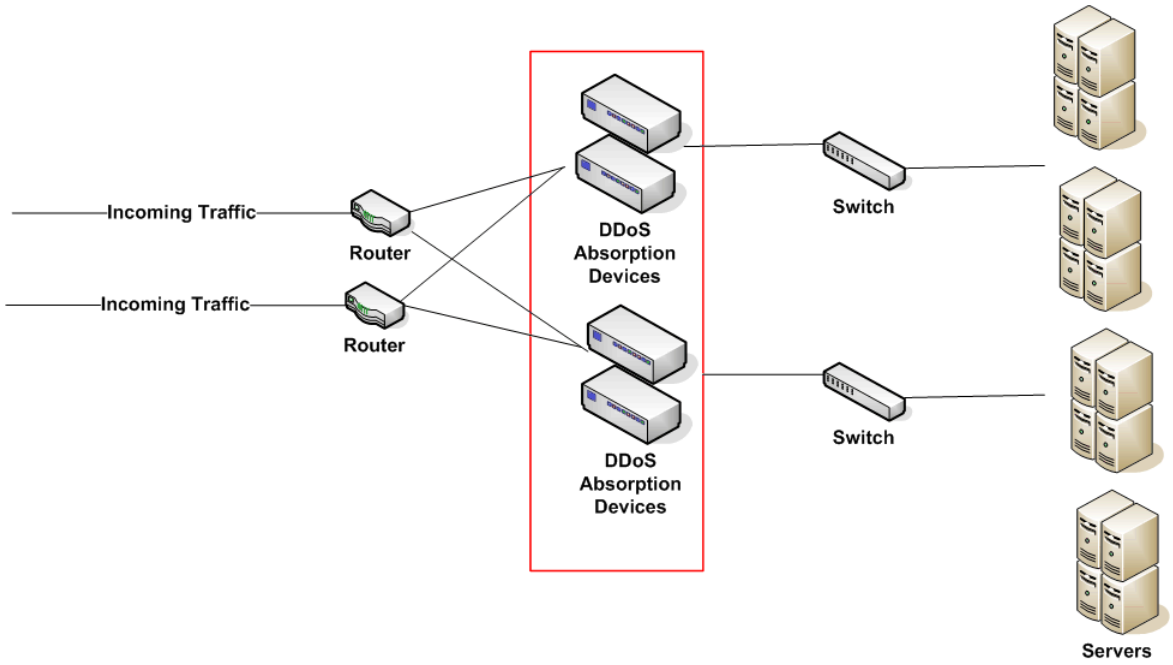

Figure 6.7: DDoS absorption blocks malicious DoS traffic before it reaches application servers.

A cloud acceleration provider should support DDoS attack mitigation. In addition to DDoS absorption devices, procedures should be in place to notify network engineers of an attack, provide detailed data about the attack, and support additional mitigation measures, such as using alternative data centers to maintain access to Web applications.

Data Security

Maintaining confidentiality of data with SSL and ensuring availability of applications with DDoS attack mitigation technologies are two key security considerations when evaluating cloud acceleration providers. A third key consideration is ensuring data security with regards to government and industry regulations.

Businesses and government agencies may be subject to multiple data protection regulations such as the Sarbanes Oxley (SOX) Act, the Health Insurance Portability and Accountability Act (HIPAA), the Payment Card Industry Data Security Standard (PCI DSS), and others. These regulations have specific requirements for protecting the confidentiality and integrity of data. Cloud acceleration providers should offer sufficient controls to allow customers to meet regulatory requirements.

Although regulations vary in their requirements, there are common characteristics such as the need for access controls, confidentiality measures, and the ability to demonstrate compliance. As part of your evaluation, verify cloud acceleration vendors are in compliance with relevant regulations for your business.

Authentication

Authentication is one area of security that sounds fairly straightforward but can be riddled with organizational challenges. Authentication is the process of verifying the identity of a user prior to granting that user access to data, applications, and systems. This process requires an authentication infrastructure that supports:

- A database of users and identifying information
- Authentication mechanisms, such as passwords or digital certificates
- Two-factor authentication, in some cases
- Life cycle management services

Content served from a content delivery network may require authentication controls. If you have such a requirement, consider how the content delivery network authenticates users and whether it meets your requirements and allows you to manage access controls with reasonable administrative overhead.

Architecture Considerations

The architecture of the content delivery network and dynamic content acceleration network is another area to assess when evaluating potential providers. Consider edge server locations, protocol optimizations, and key performance metrics.

The location of edge servers helps to shape the overall performance of your applications. Edge servers that are in close physical proximity to users will help reduce latency because packets have shorter distances to travel. Edge servers on networks with high-performance peering agreements are less likely to be subject to degraded performance when using other ISPs' networks.

Protocol optimizations are especially important for dynamic content acceleration. TCP has changed over time to support several types of optimizations that can improve throughput. These optimizations can benefit both static and dynamic content because they are typically applied to network traffic between edge servers.

Together, the location of edge servers and protocol optimizations can improve the overall performance of your applications. To quantify those improvements, look to multiple key performance metrics:

- Reduced latency or load time
- Reduced HTTP request failures
- Improved throughput
- Reduced origin load

The provider should offer analysis and reporting tools that make these and other key performance indicators readily available. In the event that performance degrades or you experience other problems—for example, disruption in encryption services because of a problem with an SSL certificate—the provider should be in a position to offer support 24x7.

Key Business Considerations

A cloud application acceleration provider will become a partner in delivering access to your applications and data. Business considerations should be evaluated along with technical considerations. Key business areas include:

- Cost—Consider startup costs, such as design, consulting, and new equipment, as well as ongoing operational costs. Also consider any costs associated with changing services or configuration.
- Reliability—Failures in content delivery network or dynamic content delivery services could lead to failures in delivering your services. Consider providers' past performance, service level agreements (SLAs), and compensation for downtime. Also carefully review how downtime is calculated and the process for submitting claims.
- Scalability—One of the advantages of cloud computing is the ability to rapidly scale up the number of servers used to deliver application services. If, however, the network cannot scale as well, the benefits of cloud scalability can be undermined.
- Security—The availability of applications and the integrity and confidentiality of your data are key security considerations.
- User experience—How is the user experience altered by deploying your application through a content delivery network and accelerated network? Besides reduced latency, are there other changes to the way applications perform that could substantially alter the user experience?

- Deployment time—How quickly can you reach new geographic regions or increase scalability in a region?
- Management support—How does the provider support your management of the network? Is technical support available 24x7?
- Analysis and reporting—What types of reports are available to help you manage ongoing operations? Are reports sufficient to support compliance with regulations?

Ideally, your provider will be able to assume the role of expert for managing implementation details of the content delivery network. In some cases, they can also become intermediaries dealing with government regulations. Of course, these technical considerations are all designed to support core business requirements related to meeting customer needs and expectations.

Impact of Slow Applications

Businesses deploying Web applications are facing a combination of two factors that could adversely affect their application performance: increasingly complex Web sites and a geographically distributed customer base.

Adverse Customer Experiences

Web designers are making use of development tools and techniques designed to improve a user's experience. Features that were once restricted to desktop applications are now available in Web applications. This improvement is welcome but it often comes at the cost of increased application complexity. Complexity, in turn, can lead to longer page load times. Research indicates that users expect a response within about 3 seconds. Sites that experience longer load times can anticipate users will abandon the site at rates significantly higher than those sites that maintain a sub-3 second average response time.

In addition, expanding your customer base into new geographic areas should be a positive experience, but high latency in Web applications can undermine your ability to deliver a suitable customer experience. Poor page loading performance can also impose a drag on innovation. Developers and designers may be hesitant to add features that enhance a user experience but would further prolong page load times. Lack of innovation over time can lead to sites that appear dated and lacking functionality. At the same time you are constrained in your options because of poor application performance, competitors may be adding features such as real-time inventory lookup, videos, optimized content, and interactive features. Baseline expectations of Web applications are constantly evolving, and high latency and other performance issues can inhibit your ability to continue to meet those changing expectations.

Ultimately, poor application performance translates into lower revenues. For example, a study by the Aberdeen Group found that a 1-second delay in page loading led to an 11% drop in page views and a 7% loss in sales. Even when customers finish a transaction on a poorly performing site, their chances of returning drop significantly. 79% of customers are less likely to purchase from a vendor in the future if they are dissatisfied with the Web site's performance.

Accelerating application performance allows application designers to continue to innovate and deliver quality user experiences. Just as important, it provides the means to maintain performance required to reduce the risk that customers will abandon shopping carts, switch to competitor sites, or otherwise abandon an application or site.

Summary

Delivering applications to a global user base is challenging. You will face technical difficulties as well as cultural issues. Business needs are driving the adoption of cloud acceleration to improve the overall performance of applications. Technical considerations are best addressed with a combination of data centers, content delivery networks, and dynamic content acceleration techniques. As this chapter has outlined, there are multiple considerations to evaluate when assessing content delivery network providers. The importance of particular considerations will vary according to your specific business requirements. Considering the full range of technical, business, and cultural issues you face in delivering content to a global user base will help you evaluate your content delivery network and cloud application acceleration options.