Choosing a Content Delivery Method

Executive Summary

Cache-based content distribution networks (CDNs) reach very large volumes of highly dispersed end users by duplicating centrally hosted video, audio and data content across the Internet. Geographically distributed cache servers reduce the proximity between user and content, thereby reducing latency and increasing the performance of real-time traffic.

Content owners wishing to widely distribute their content, but not in a position to build and manage a worldwide cache CDN themselves, have two primary sources for procuring cache-based CDN services:

1. a Network-based CDN provider that also offers WAN and Internet infrastructure/access connectivity services to end users, or

2. a CDN provider that installs and operates the CDN cache overlay to the network infrastructure.

As this paper discusses, there are distinct advantages of turning to a Network-based CDN provider, who is responsible for the basic routing, troubleshooting, operations and management of the network at Layers 1-3 in addition to providing the CDN services. By contrast, the CDN-only provider typically has no visibility into Layer 1-3 network conditions and thus has comparatively limited control over content delivery performance.

at&t

Introduction

News organizations, entertainment companies, educational institutions and a host of other parties need to reliably deliver multimedia content to millions of Internet-connected individuals. For example, they might like to stream live or recorded sporting events, movies, seminars, newscasts and fashion shows to users' computers around the world. A type of network called a content delivery network (CDN), which is basically a specialized server overlay to the public Internet, has been specifically designed and optimized to deliver such high-bandwidth Internet content reliably to large volumes of widely distributed recipients.

Types of CDNs

Content can be distributed across the Internet using several different CDN mechanisms. Among them:

- **Unicast. Centralized servers hosting the original content serve user requests. This method works well for low-demand content or content that is customized to the end user.**
- **Cache-Based. Content nodes distributed across the Internet serve user requests. The node physically closest to the end user request serves that request, minimizing the content travel distance. Cache CDNs can be application-based only or both application-based and network-based. Cache CDNs work well for high-demand content.**
- **Peer-to-Peer. P2P CDNs deliver content from an origin server initially. Once there are many copies distributed to user PCs, the PCs can take over as content nodes in their own right and deliver the content directly to other PCs.**
- **Multicast. Still in a nascent phase, multicast CDNs send content simultaneously to end users by adaptively replicating and branching streams of data across the IP network to deliver content only to those end users that request it. All users share a single data stream for network efficiency. Multicast will be a strong option for distributing live streaming applications and any content in high demand.**

There are several kinds of CDNs, though some are in more mature development phases than others (see box above). Today, cache-based CDNs deliver the highest-quality and most reliable experience to very large volumes of dispersed end users. The reason is that they operate by duplicating the video, audio and data content hosted on centralized origin servers, where the original content is stored. Cache CDNs store copies in Internet-connected servers, also called content nodes or cache servers, which are distributed globally. As such, there is a copy, or cache, of the content geographically close to most users.

The cache CDN setup reduces the content delivery delay, or latency, imposed by the long physical distances over which content would otherwise have to travel to users who might be half a world away from the central origin servers. Too much latency and its cousin, jitter, or the variation in the amount of latency, degrade the quality of live and streaming video and audio by making it appear jerky or sound warbled. So, reducing latency and jitter by shortening the distance between users and content servers is necessary for delivering good-quality and consistent user experiences.

Build or Buy?

There are several approaches to using a cache CDN to distribute content. It's theoretically possible for the content owners to build and manage the CDN themselves. This strategy is labor-intensive, however, and often proves impractical. Many content providers' primary business focus is either on creating the content itself or on other endeavors of which the content is a key part. Building and maintaining a worldwide network of content nodes requires substantial capital investments as well as expertise in constructing, maintaining and securing large-scale CDN server networks. This represents a level of complexity and cost likely to distract businesses from their core competencies.

There are also specialized CDN providers who offer the server overlay capabilities in the form of a service. These independent companies connect and manage a given content provider's content nodes all over the Internet. They can control server congestion and content delivery response times to the degree that the part of the network at which they operate – Layers 4 through 7 – allows. They are focused on content delivery as their primary business; however, they lack direct visibility into the underlying network foundation that interconnects their servers. Rather, they send probes across the network to gather application-layer information that they have to interpret – to varying degrees of accuracy – before taking action to circumvent congestion and outages. Without knowledge of changing network conditions, they are challenged to optimize content delivery performance.

A third, often most advantageous, choice for high-demand content delivery is to use a network service provider that also offers the specialized CDN cache service as a managed service. The network-based provider not only has the application-layer congestion, routing and management tools of the specialized CDN provider; it also has all the Layer 3 network-based IP visibility, routing information and associated intelligent services needed to optimize underlying network efficiency.

The remainder of this paper compares using the cache-based content delivery services of a large IP network provider to a specialized CDN service supplier.

Application-Layer vs. Network-based CDNs

Most cache-based CDNs today are operated by independent CDN providers and work at the application layer only. Content nodes are installed and managed by a CDN service provider that procures the IP network routing services that actually transport content from a third-party network service provider. The application-based CDN service provider thus has little or no knowledge of the underlying IP network; rather, the contracted network service provider manages and secures the routed network independently of the needs of the CDN.

In an application-layer CDN, the content nodes use Domain Name System (DNS) information to approximate an end user's location. In other words, the application-layer CDN redirects user requests for content from the DNS server to which the user initially connects to the content node nearest that DNS server. DNS is required for basic Web

browsing, as it resolves hostnames to actual IP addresses. This redirection may or may not be optimal, however, because user connections to DNS servers aren't geographically based. So the DNS server might not be anywhere near the content node that is physically closest to the user.

Network-based CDNs delivered as a managed service option by a network service provider, by contrast, combine visibility of the DNS application-layer information with that of the IP network layer, including routing tables, router status, network congestion and Layer 2 connection types and status. Network-based CDNs also have the ability to use IP-layer intelligent services such as traffic engineering to closely manage network conditions and optimize content delivery.

In addition, a Layer 3 capability called IP Anycast – an Internet Engineering Task Force (IETF) standard – is often deployed in network-based CDNs to allow the same IP address to be assigned to multiple content nodes. This allows user requests to be served by any of a number of nodes based on user location and the state of network conditions at the time of the request. Of those nodes with the same IP address, the network routing protocol determines the best and closest content node to the user – something the application-layer CDN alone cannot do.

Historically, to determine the "best" node for redirection of users, IP Anycast has used traditional routing conditions, including shortest path, network congestion, router and link status and least-cost route information. However, some network service provider implementations of IP Anycast are also gaining an adaptive element, so that the redirection of user content requests also accounts for the load on a node at the time of the request. In this way, when selecting among multiple nodes with the same IP Anycast address, the network can consider both how busy the server is at the moment, or its load, and the network conditions described to deliver the optimal experience to the end user.

Network-based CDNs, then, have distinct operational and efficiency advantages over application-based CDNs for maximizing the quality of the end user experience (see Table 1).

The Power of Network Ownership

A network service provider owns many important assets that are critical to high-quality content delivery. Major network providers typically offer a global network reach and a large base of direct enterprise and consumer access links. They enjoy strong peering relationships with most other network providers, further extending their coverage.

The network provider also has a steady stream of network performance data coming into a network operations center (NOC) with technicians prepared to react to any issues on a 24/7 basis. Security tools monitor traffic flows and security experts are ready to act on any issues detected on the network. When the network provider also offers a content delivery service, these same capabilities become assets of the CDN.

Table 1: CDN Operational Comparison

Function	Application-Layer CDN	Network-Based CDN
Cache Server Selection	Routes to cache servers based upon the location of an end user's DNS	Routes to most efficient cache server based upon the actual source location of the request, minimizing content transmission distance and latency
Troubleshooting	Detects congestion and outages based on indirect probing and performance measures	Detects congestion and outages directly from IP routing and link monitoring information, accelerating resolution time
Rerouting around Outages	Reroutes around detected outages based upon relative performance of alternates	Reroutes around detected outages based upon actual, dynamic network conditions in addition to the relative performance of alternates, accelerating delivery
Rerouting around Congestion	Reroutes around detected congestion based upon the relative performance of alternates	Dynamically changes IP routing to eliminate congestion within the provider's own IP backbone and peer IP networks
Content Node Location	CDN provider locates content nodes in third-party data centers. User requests are aggregated in a local switching center first, and then forwarded to the data center for service, inducing latency	Network provider collocates content nodes in major routing centers, where user requests are handled locally
Load Balancing	Distributes content requests based upon the resources available in the cache servers. Load balancing might be slowed by the time it takes for DNS changes to propagate across the Internet (often 10 or more minutes)	Distributes content requests based upon the resources available in the cache servers, as well as in access routers, core routers and peering links, for a more informed and better-performing decision. Load balancing occurs in the time it takes for routing updates to propagate the Layer 3 network (milliseconds)
Knowledge/Ownership of End Customer	Has no direct knowledge of end user location; uses DNS information only. DNS server could be in one state, while the subscribing user could be in another	Has direct knowledge of its consumer and business network customers by IP address, improving the proximity of cache server to user

In addition, the provider gives network access and other ancillary services to the very same end users who want to consume the content. Because it owns the user access infrastructure, the routing infrastructure and the content distribution/caching infrastructure – and it has operational responsibility for all those components – the network provider is in the best position to manage the content delivery experience of the end customer. The network provider, in effect, serves two masters: the owner of the content requiring network distribution and the end user who consumes both the content and network-connectivity services.

The network provider has access to the physical network facilities throughout its network. This access can enable greater optimization in the placement of CDN equipment. For example, it can place equipment in network nodes close to large groups of its end users and collocate nodes in network peering locations. These locations further minimize distance and latency for content distribution.

As a result of owning the content distribution, routing and access infrastructure, the network provider is directly aware of details of the routing, performance, congestion and failures over the entire path from the source of the content to its destination of end user computers. It also has knowledge of the time, origin and best destination of the end user request for content in the form of the DNS query. This allows the network provider to know and directly measure the impact of its performance on the end user and to be most efficient in its proximity routing. Independent application-layer CDN providers that do not own end user access are aware only of the network that generated the user request, not the details of the routing all the way to the end user. As a result, their proximity routing is inherently less efficient.

The network-based service provider can respond quickly to correct any content delivery issues because it receives notice of failures and congestion directly from the monitored network infrastructure. It does not need to rely on third parties for troubleshooting information or wait for an analysis of test packets sent into various networks. An application-layer CDN provider would need to respond based on only indirect information about the relationship between the end users and network traffic flows.

The network provider has a direct relationship with its end users and is frequently in contact with them through service requests, billing, trouble resolution and sales efforts. The network-based provider, then, can enhance the content provider/end user customer relationship and provide the content owner with an additional means of contact with end users. The two sources functioning together can have a more effective impact on end users than either one alone. For example, the network-based CDN provider can offer the content provider the ability to use space on its portal or other means of end user communications to notify consumers of the availability of new content. The content provider would otherwise be able to reach only end users who affirmatively visit its Web site, and it could not autonomously direct targeted information to specific groups of end users.

Likewise, the CDN customer can use a single source for its network services and content delivery. It has access to multiple services through a single portal and a single provider to maintain, troubleshoot and manage the network.

The Double-Duty Advantage

It is beneficial to both providers of content and content consumers for the same entity to operate the network infrastructure and the CDN service overlay. An ICDS provider that owns the network down to the end user:

- **Has responsibility for and control over a wide set of resources used to deliver content to and from enterprises, consumers and peer networks**
- **Can optimize the placement of content delivery equipment to be physically close to end users**
- **Is aware of routing details over the full path to the end user**
- **Has direct knowledge of congestion and/or failures all the way to the end user**
- **Has a direct interest in satisfying both the content provider customer and the end user customer**
- **Receives and provides feedback directly to and from end users**
- **Can directly support and enhance the relationship between the content provider and the end user**

Using Routing Intelligence to Optimize Delivery

A network service provider that doubles as a CDN provider has the opportunity to blend network information with content distribution information to achieve higher levels of content delivery stability and performance. This level of service and network integration is possible, in part, with a new technology developed by AT&T called Intelligent Route Service Control Point (IRSCP). IRSCP extends traditional network route selection capabilities to account for application performance conditions.

For example, node congestion is a common problem for CDN providers because demand is unpredictable and subject to spikes. An IRSCP application, which creates a direct relationship between the CDN and network routing controls, can provide the network with advanced load-distribution capabilities in such situations. A network-based, IP Anycast-based CDN service provider using IRSCP can adapt the IP Anycast route to steer traffic away from a busy node to other nodes before congestion occurs. Content node CPU utilization, content demand information, network link performance and other network operational information can be continuously fed into an IRSCP application to intelligently balance content requests across nodes.

The IRSCP application receives real-time statistics from the network and CDN service. As such, it doesn't depend on sending test probes across the network to determine the performance of the network; rather, the adjustment to traffic flows is immediate. As soon as IRSCP issues a route update to the routers, a portion of the traffic begins to flow away from the nearly congested node. This avoids the delay

experienced while application-layer CDN providers wait for DNS cache refreshes to ripple through the network. Since application-layer CDN providers lack the IP route visibility and IRSCP control, they are unable to achieve this optimization.

Preserving Large File Downloads

Large file downloads have become very popular for CDNs and can last for several minutes. When CDN providers make adjustments to traffic flows, they need to avoid breaking any large file downloads in progress. In a typical IP Anycast implementation, a sudden routing change for the Anycast address would disrupt large file download sessions in progress and cause users to restart the download. This undesirable situation can be avoided by combining IP Anycast and traditional unicast routing technologies: A large file download request is switched from the IP Anycast address to a unicast IP address to preserve the download session if the IP Anycast address routing is altered.

Integrated Network and Application Service Security

Integrating network security and content-delivery service security creates defense in depth. For its part, the IP network detects, isolates and eliminates most threats before they become breaches. For example, it monitors traffic flows for early warning signs of new distributed denial of service (DDoS) attacks such as those caused by worms and viruses. If data indicate that a potential attack is brewing, the network security team can be alerted to take proactive steps to minimize any adverse effect on the network and CDN service assets.

DDoS attacks attempt to disrupt service by flooding a CDN with a large number of packets to exhaust available resources. If the CDN nodes are deployed within the provider's own network, the provider can implement routing updates to divert the DDoS attack traffic away from the CDN service. An effective DDoS prevention service can isolate the DDoS attack traffic from valid customer traffic so that legitimate customer requests continue to be serviced by the CDN.

Other security mechanisms also help protect against DDoS attacks. For example, CDN security rules only allow access to specific services from specific IP addresses. Configuration settings within firewalls, routers, switches and servers minimize the impact of packet flooding. Regular and frequent audits of all network elements help ensure that their security settings are up-to-date. Since the security threat response team protects the overall IP network, it has access to more and better information about the nature of the security threat. This is particularly important for maintaining service in the presence of an elevated number of requests.

IP network service providers require strict security on the network equipment that runs the network. Network switching facilities are typically deployed in very secure physical facilities with highly reliable power and cooling. Access to the network equipment is restricted to authorized personnel only. A combined network and CDN service provider will extend this high level of physical security, access security and facility reliability to the CDN service. In this scenario, any threats against the CDN node receive the same level of response as if it were a threat against the IP network.

Summary

A content delivery network is optimized to deliver high-bandwidth, latency-sensitive traffic such as streaming video and audio to Internet users. To deliver high-demand content to widely distributed Internet users, a cache-based CDN that stores copies of content out near user access points delivers the best performance, because it reduces distance-imposed latency and jitter, which can disrupt video and audio quality and, thus, a user's experience.

There are two types of cache-based CDNs: application-layer-only and network-based CDNs. Network-based CDN services are available from providers who own and operate a network infrastructure. By being network-based, they combine DNS application-layer information about user location and Layer 3 IP control information about end user routing location and network conditions. Armed with routing control plus DNS visibility, network-based CDN service providers exercise tight control over end user content experiences. By contrast, specialty companies that offer application-layer-only CDN services are not able to tell as much about network conditions or server congestion in routing user content requests.

In network-based CDNs, new IRSCP applications, used in conjunction with traditional IP Anycast address broadcasting, account for how busy a given content node is when redirecting user content requests. Both content providers and content consumers benefit by using a single network service provider that owns infrastructure and the NOC, because end users benefit from DNS, routing and server congestion management information being combined to deliver the ultimate user experience.

For more information contact an AT&T Representative or visit www.att.com/business.

07/02/08 AB-1387
© 2008 AT&T Intellectual Property. All rights reserved. AT&T and the AT&T logo are trademarks of AT&T Intellectual Property.

at&t
Your world. Delivered.