Globally Distributed Content Delivery

Using more than 12,000 servers in over 1,000 networks, Akamai's distributed content delivery system fights service bottlenecks and shutdowns by delivering content from the Internet's edge.

John Dilley, Bruce Maggs, Jay Parikh, Harald Prokop, Ramesh Sitaraman, and Bill Weihl
Akamai Technologies

As Web sites become popular, they're increasingly vulnerable to the flash crowd problem, in which request load overwhelms some aspect of the site's infrastructure, such as the front-end Web server, network equipment, or bandwidth, or (in more advanced sites) the back-end transaction-processing infrastructure. The resulting overload can crash a site or cause unusually high response times — both of which can translate into lost revenue or negative customer attitudes toward a product or brand.

Our company, Akamai Technologies, evolved out of an MIT research effort aimed at solving the flash crowd problem (www.akamai.com/en/html/about/history.html). Our approach is based on the observation that serving Web content from a single location can present serious problems for site scalability, reliability, and performance. We thus devised a system to serve requests from a variable number of surrogate origin servers at the network edge.[1] By caching content at the Internet's edge, we reduce demand on the site's infrastructure and provide faster service for users, whose content comes from nearby servers.

When we launched the Akamai system in early 1999, it initially delivered only Web objects (images and documents). It has since evolved to distribute dynamically generated pages and even applications to the network's edge, providing customers with on-demand bandwidth and computing capacity. This reduces content providers' infrastructure requirements, and lets them deploy or expand services more quickly and easily. Our current system has more than 12,000 servers in over 1,000 networks. Operating servers in many locations poses many technical challenges, including how to direct user requests to appropriate servers, how to handle failures, how to monitor and control the servers, and how to update software across the sys-

tem. Here, we describe our system and how we've managed these challenges.

Existing Approaches

Researchers have explored several approaches to delivering content in a scalable and reliable way. Local clustering can improve fault-tolerance and scalability. If the data center or the ISP providing connectivity fails, however, the entire cluster is inaccessible to users. To solve this problem, sites can offer mirroring (deploying clusters in a few locations) and multihoming (using multiple ISPs to connect to the Internet). Clustering, mirroring, and multihoming are common approaches for sites with stringent reliability and scalability needs. These methods do not solve all connectivity problems, however, and they do introduce new ones:

- It is difficult to scale clusters to thousands of servers.
- With multihoming, the underlying network protocols — in particular the border gateway protocol (BGP)[2] — do not converge quickly to new routes when connections fail.
- Mirroring requires synchronizing the site among the mirrors, which can be difficult.

In all three cases, excess capacity is required: With clustering, there must be enough servers at each location to handle peak loads (which can be an order of magnitude above average loads); with multihoming, each connection must be able to carry all the traffic; and with mirroring, each mirror must be able to carry the entire load. Each of these solutions thus entails considerable cost, which could more than double a site's initial infrastructure expense and ongoing operation costs.

The Internet is a complex fabric of networks. Congestion and failures occur at many places, including

- the "first mile" (which is partially addressed by multihoming the origin server),
- the backbones,
- peering points between network service providers, and
- the "last mile" to the user.

Deploying independent proxy caches throughout the Internet can address some of these bottlenecks. Transit ISPs and end-user organizations have installed proxy caches to reduce latency and bandwidth requirements by serving users directly from a previously requested content cache. However, Web proxy cache hit rates tend to be low — 25 to 40 percent — in part because Web sites are using more dynamic content. As a result, proxy caches have had limited success in improving Web sites' scalability, reliability, and performance.

Akamai works closely with content providers to develop features that improve service for their Web sites and to deliver more content from the network edge. For example, features such as authorization, control over content invalidation, and dynamic content assembly let us deliver content that would otherwise be uncacheable. Although ISP caches could include similar features, to be useful they would have to standardize the features and their implementation across most cache vendors and deployments. Until such a feature is widely deployed, content providers have little incentive to use it. Because Akamai controls both its network and software, we can develop and deploy features quickly.

Akamai's Network Infrastructure

Akamai's infrastructure handles flash crowds by allocating more servers to sites experiencing high load, while serving all clients from nearby servers. The system directs client requests to the nearest available server likely to have the requested content. It determines this as follows:

- *Nearest* is a function of network topology and dynamic link characteristics: A server with a lower round-trip time is considered nearer than one with a higher round-trip time. Likewise, a server with low packet loss to the client is nearer than one with high packet loss.
- *Available* is a function of load and network bandwidth: A server carrying too much load or a data center serving near its bandwidth capacity is unavailable to serve more clients.
- *Likely* is a function of which servers carry the content for each customer in a data center: If all servers served all the content — by round-robin DNS, for example — then the servers' disk and memory resources would be consumed by the most popular set of objects.

In the latter case, an Akamai site might hold a dozen or more servers within any data center; the system distributes content to the minimum number of servers at each site to maximize system resources within the site.

Automatic Network Control

The direction of requests to content servers is referred to as mapping. Akamai's mapping tech-

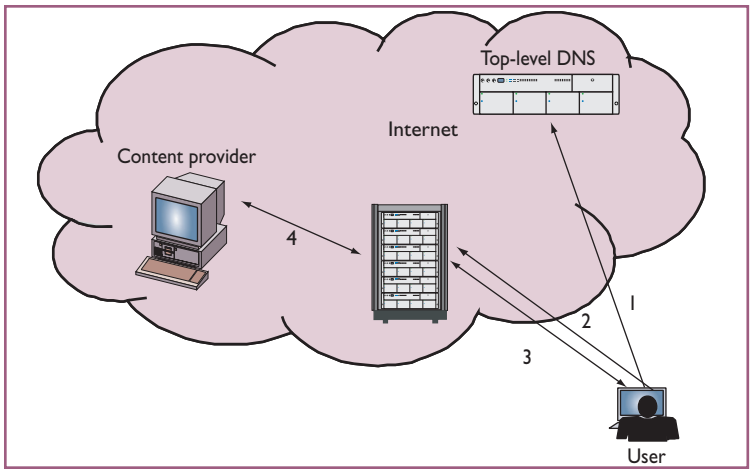

Figure 1. Client HTTP content request. Once DNS resolves the edge server's name (steps 1 and 2), the client request is issued to the edge server (step 3), which then requests content (if necessary) from the content provider's server, satisfies the request, and logs its completion.

nology uses a dynamic, fault-tolerant DNS system. The mapping system resolves a hostname based on the service requested, user location, and network status; it also uses DNS for network load-balancing.

The "DNS Resolution" sidebar describes the standard DNS resolution process for an Akamai edge server name, corresponding to steps 1 and 2 in Figure 1. In step 3, the client makes an HTTP request to the edge server, which then retrieves the content by requesting it from either another Akamai server or the content provider's server (step 4). The server then returns the requested information to the client and logs the request's completion.

Akamai name servers resolve host names to IP addresses by mapping requests to a server using some or all of the following criteria.

- *Service requested*. The server must be able to satisfy the request. The name server must not direct a request for a QuickTime media stream to a server that handles only HTTP.
- *Server health*. The content server must be up and running without errors.
- *Server load*. The server must operate under a certain load threshold and thus be available for additional requests. The load measure typically includes the target server's CPU, disk, and network utilization.
- *Network condition*. The client must be able to reach the server with minimal packet loss, and the server's data center must have sufficient bandwidth to handle additional network requests.
- *Client location*. The server must be close to the client in terms of measures such as network round trip time.
- *Content requested*. The server must be likely to have the content, according to Akamai's consistent hashing algorithm.

Internet routers use BGP messages to exchange network reachability information among BGP systems and compute the best routing path among the Internet's autonomous systems.[2] Akamai agents communicate with certain border routers as peers; the mapping system uses the resulting BGP information to determine network topology. The number of hops between autonomous systems is a coarse but useful measure of network distance. The mapping system combines this information with live network statistics — such as traceroute data[3] — to provide a detailed, dynamic view of network structure and quality measures for different mappings. Implementing this mapping system on a global scale involves several challenges, as we discuss later.

Network Monitoring

Our DNS-based load balancing system continuously monitors the state of services, and their servers and networks. Each of the content servers — for the HTTP, HTTPS, and streaming protocols — frequently reports its load to a monitoring application, which aggregates and publishes load reports to the local DNS server. That DNS server then determines which IP addresses (two or more) to return when resolving DNS names. If a server's load exceeds a certain threshold, the DNS server simultaneously assigns some of the server's allocated content to additional servers. If the load exceeds another threshold, the server's IP address is no longer available to clients. The server can thus shed a fraction of its load when it is experiencing moderate to high load. The monitoring system also transmits data center load to the top-level DNS resolver to direct traffic away from overloaded data centers.

To monitor the entire system's health end-to-end, Akamai uses agents that simulate end-user behavior by downloading Web objects and measuring their failure rates and download times. Akamai uses this information to monitor overall system performance and to automatically detect and suspend problematic data centers or servers.

In addition to load-balancing metrics, Akamai's monitoring system provides centralized reporting on content service for each customer and content server. This information is the basis of Akamai's real-time customer traffic analyzer application. The information is useful for network operational and diagnostic purposes, and provides real-time

DNS Resolution

Akamai edge servers are located using a DNS name, such as a7.g.akamai.net. A DNS resolver resolves this name in the standard manner, from right to left, querying DNS name servers until IP addresses for the host a7 in the domain .g.akamai.net are returned.

Each name resolution associates a "time to live" (TTL) with the resolution, which proceeds as follows:

1. The resolver chooses a root name server and asks it to resolve the name a7.g.akamai.net. The root name server does not itself resolve the name; instead, it sends a domain delegation response with IP addresses of the name servers that handle .net domain requests.
2. The resolver then queries the .net name servers, which return a domain delegation (NS records) for .akamai.net. These are the Akamai top-level name servers (top-level DNS in Figure 1).
3. Next, the resolver queries an Akamai TL DNS server, which returns a domain delegation for .g.akamai.net to low-level Akamai name servers (low-level DNS in Figure 1) with a TTL of about one hour. The low-level name servers selected correspond to (and are in the same location as) the available edge servers that are closest to the requesting user.
4. Finally, the resolver queries an Akamai low-level DNS server, which returns the IP addresses of servers available to satisfy the request. This resolution has a short TTL (several seconds to one minute), which encourages frequent refreshes of the DNS resolution and allows Akamai to direct requests to other locations or servers as conditions change.

A resolver is preconfigured to know the Internet root name servers' IP addresses, which are the starting points for a DNS resolution if the resolver lacks required information in its cache. If the resolver has valid IP addresses of the .net name server, it skips step 1; if it has cached IP addresses of the g.akamai.net name servers, it skips steps 1 through 3.

The resolution process is the same for any DNS name. Akamai name resolution differs, however, in how its name servers behave.

access to an array of service parameters organized as a database. The application's SQL-like interface supports ad hoc queries against live and historic data, which lets the operations staff locate the busiest customer, the server using the most memory or disk space, or the switch or data center closest to its bandwidth limit.

Network Services

Akamai servers deliver several types of content: static and dynamic content over HTTP and HTTPS, and streaming audio and video over the three streaming protocols described below.

Static Content

Static Web content consists of HTML pages, embedded images, executables, PDF documents, and so on. Akamai's content servers use content type to apply lifetime and other features to static documents, which have varying cacheability and can have special service requirements.

Lifetimes, for example, can vary from zero seconds, where the edge server validates the object with the origin server on each request, to infinite, where the content server never checks the object's consistency. Lifetime values for Akamai edge servers can also differ from downstream proxy servers and end users.

Special features might include the ability to serve secure content over the HTTPS protocol, support alternate content and transfer encodings, handle cookies, and so on. Akamai controls features on behalf of each customer using a metadata facility that describes which features to apply by customer, content type, and other criteria.

Dynamic Content

Today's Web sites depend heavily on dynamic content generation to offer end users rich and captivating material. As we noted earlier, however, proxy caches cannot typically cache dynamic content. A proxy cache could not, for example, handle a largely static Web page if it contained an advertisement that changed according to each user's profile.

To deal with this, we use Edge Side Includes technology (www.esi.org), which assembles dynamic content on edge servers. ESI is similar to server-side include languages, but adds fault-tolerance features (for when the origin server is unavailable) and integrates an Extensible Stylesheet Language Transformation (XSLT) engine to process XML data. Using ESI lets a content provider break a dynamic page into fragments with independent cacheability properties. These fragments are maintained as separate objects in the edge server's cache and are dynamically assembled into Web pages in response to user requests.

The ability to assemble dynamic pages from individual page fragments means that the server must fetch only noncacheable or expired fragments from the origin Web site; this reduces both the load on the site's content generation infrastructure and the data that the edge server must retrieve from the central origin server. ESI reduced bandwidth

requirements for dynamic content by 95 to 99 percent across a range of dynamic sites we studied, including portals and financial sites. The resulting reduction in central infrastructure offers content providers significant savings.

Streaming Media

Akamai's streaming network supports live and on-demand media in the three major formats — Microsoft Windows Media, Real, and Apple's QuickTime. While building a streaming delivery network presents some technical issues that are similar to those of a Web delivery network, there are also significant additional challenges.

First, the content provider typically captures and encodes a live stream and sends it to an entry-point server in the Akamai network. Given our principle of removing all single points of failure, we must have mechanisms that will react quickly to a failed entry-point server. Specifically, another entry-point server must pick up the live stream quickly enough that end users detect no interruption in the stream.

The stream is delivered from the entry-point server to multiple edge servers, which in turn serve the content to end users. Media packet delivery from the entry-point to the edge servers must be resilient to network failures and packet loss, and thus the entry point server must route packet flows around congested and failed links to reach the edge server. Further, the entry point and edge servers must deliver packets without significant delay or jitter because a late or out-of-order packet is useless in the playback. When necessary, Akamai uses information dispersal techniques that let the entry point server send data on multiple redundant paths, which lets the edge server construct a clean copy of the stream even when some paths are down or lossy.

Typically, a content provider uploads an on-demand clip into an Akamai content storage facility. We distribute the storage facility over many data centers and automatically replicate the uploaded clip to a subset of the centers. An edge server that receives a stream request downloads the content from its optimal storage location and caches it locally while serving the request.

Technical Challenges

Constructing a global network poses many non-technical challenges, including deploying network equipment and server hardware, establishing good working relationships with network providers, controlling operational expenses, and acquiring and supporting customers. While these challenges are significant, our focus here is on challenges related to designing, building, and operating the system itself.

System Scalability

Akamai's network must scale to support many geographically distributed servers, and many customers with differing needs. This presents the following challenges.

- Monitoring and controlling tens of thousands of widely distributed servers, while keeping monitoring bandwidth to a minimum.
- Monitoring network conditions across and between thousands of locations, aggregating that information, and using it to generate new maps every few seconds. Success here depends on minimizing the overhead added to DNS to avoid long DNS lookup times. This lets us perform the calculations required to identify the optimal server off-line, rather than making the user wait.
- Dealing gracefully with incomplete and out-of-date information. This requires careful design and iterative algorithm tuning.
- Reacting quickly to changing network conditions and changing workloads.
- Measuring Internet conditions at a fine enough granularity to attain high-probability estimates of end-user performance.
- Managing, provisioning, and solving problems for numerous customers with varying needs, varying workloads, and varying amounts of content.
- Isolating customers so they cannot negatively affect each other.
- Ensuring data integrity over many terabytes of storage across the network. Because low-level (file system or disk) checks are inadequate to protect against possible errors — including those caused by operators and software bugs — we also perform end-to-end checks.
- Collecting logs with information about user requests, processing these logs (for billing), and delivering accurate, timely billing information to customers.

To meet the challenges of monitoring and controlling content servers, Akamai developed a distributed monitoring service that is resilient to temporary loss of information. To solve problems for customers, Akamai has customer-focused teams that diagnose problems and provide billing services.

System Reliability

A distributed system offers many opportunities for

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.