Akamai's EdgePlatform for Application Acceleration

Transform the Internet into a Business-Ready Application Delivery Platform— Fast Enterprise Applications, No Hardware Required

Table of Contents

Overview

Enterprise applications are increasingly becoming "Webified" to leverage the reach and cost efficiencies of the Internet to extend business processes globally. The enterprise has grown rapidly to deploy Web-based personal productivity and collaboration applications like Microsoft® Outlook® and SharePoint®, enterprise business processes supported by SAP® and Oracle®, extranet portals and content management systems, and Web services deployments to leverage machine-to-machine interactions, to name just a few.

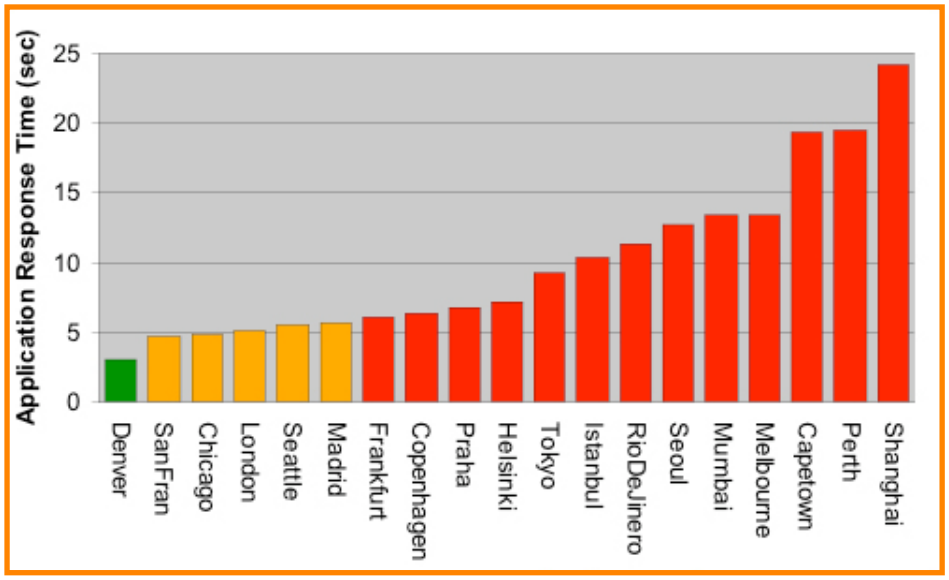
Companies face conflicting requirements when moving business critical applications and processes to the Internet—minimize the cost and management complexity associated with data center deployments while ensuring the greatest application performance and security. In addition, regulatory compliance considerations such as Sarbanes-Oxley are pressing the drive to centralize application servers into as few locations as possible.

To the surprise of many who have either developed in-house or purchased off-the-shelf Web-based enterprise applications, the performance and availability of the application often fall short of the enterprise needs from an end-user perspective. In most instances, the application works well providing sub-second response times for those users in close proximity to where the application server resides. However, as users get farther away from the server, the performance deteriorates rapidly. In fact, Web application response times for users far away from the server can often exceed 10 times that for those nearby—making the application experience very poor or downright unusable for a globally distributed user base.

Middle-Mile Bottlenecks

Good performance at data center degrades with distance

- The application runs fine for users very close to the data center.
- "Middle-Mile" bottlenecks degrade performance as distance increases from user to the data center.
- Caused by inefficiencies in routing, communication and chatty application protocols



The underlying cause of the performance degradation is that the Internet's routing protocols—BGP, the transport protocol—TCP and the Web's communication protocol—HTTP—are simply not designed for optimal performance and scale, which becomes increasingly evident as distance increases. These are often referred to as "middle-mile" bottlenecks.

> *Response times for users far away from the server can often exceed 10 times that for those nearby.*

The need for high-performance and highly-available Web-based applications has created numerous solutions for application owners and IT managers to evaluate. In this whitepaper, we examine the root causes of slow and unpredictable application response times over the Internet and why the Akamai globally distributed EdgePlatform, the foundation for its dynamic Web and IP acceleration managed services, transcends both traditional CDNs as well as hardware application acceleration appliance solutions to amplify the strengths of the Internet while overcoming its performance and reliability weaknesses.

Web Application Delivery Challenges

A Web-enabled enterprise can cost-effectively provide corporations with tremendous reach and flexibility for their business-critical applications. However, the nature of Web applications has changed dramatically as they've become increasingly dynamic and interactive over time. Enterprises must improve the performance and scale of the two major classes of Web-enabled applications—dynamic applications and Web services.

IT organizations face implementation challenges across both of these classes of dynamic Web-enabled applications including:

- Poor application performance resulting in low adoption rates and ineffective business processes
- Inconsistent application availability resulting in high abandonment and poor end user satisfaction
- Unpredictable capacity planning due to spiky peak usage, often resulting in overwhelmed or excessively built out application infrastructure

Dynamic Applications

It is estimated that up to 75% of all Web applications is dynamically generated, driven by the need to serve up personalized content for mission-critical business processes. Dynamic content does not lend itself to benefitting from caching, a technique employed to improve the response time associated with static or frequently used content.

Web Services

The use of Web services, also known as machine-to-machine interaction, is becoming an increasingly common platform for strategic B2B applications. Data from one geographically distant Web service application becomes a required ingredient in another application's processing or presentation.

The Network and Inefficient Protocols Are The Culprit in Slow Application Delivery

Middle-Mile Bottlenecks to The Forefront

The inherent design of the Internet, being a best effort "network of networks", is in itself the root cause for application delivery problems for a global user base. The Internet was designed without Web-enabled enterprises in mind. We examine the root cause of dynamic Web application and Web service performance problems for globally distributed users, namely the "middle-mile"—i.e., wide-area network bottlenecks caused by the network itself and the inefficiencies of the protocols used for communicating.

RTT Is A Culprit in Slow Application Delivery

Information sent across a network always incurs some latency as it travels from the source to the destination, measured as IP (Internet Protocol) Round-Trip Time or RTT. This is the number of milliseconds it takes for one IP packet to travel from one location to another and for a response packet to be received along the same path. For example, sending a packet from Los Angeles to New York and back would never occur faster than ~ 30 msec, since this is the amount of time required for light to travel in a direct path across the United States. Although a fraction of RTT will always arise from speed-of-light limitations, network bottlenecks associated with the design of the Internet such as service provider peering points, congestion, blackouts and brownouts add unnecessary latency and packet loss, making each RTT less than ideal. For example, latency and packet loss characteristics can exceed 100 msec with 10% packet loss across a single geographic region such as the United States or Europe. Similarly, a single RTT can often exceed 500 msec with 20% packet loss between the United States and Asia-Pacific regions.

Also Consider The RTT Multiplier Effect

A complete Web page download requires not one, but several tens of back-and-forth round trips between a client and a server before the entire page has been loaded. This "RTT multiplier" increases for those application users residing further away from the origin application server. As a result, "middle-mile" bottlenecks are introduced not only by the RTT, for a single trip taken over the Internet, but also by the RTT multiplier associated with the number of times round trips must to be taken over the Internet to build a single Web page.

The RTT multiplier effect is due to Internet transport protocol (TCP), the core Web application protocol (HTTP), and the way these two protocols interact. By design, TCP requires roundtrips to establish each new connection, the so-called "three way handshake", and then additional round-trips to "warm" each new connection. TCP's aptly-named "slow-start" feature is designed to negotiate an agreed upon communication speed between the browser and the server. In general it takes three TCP protocol exchanges to set up a TCP connection for communications and four exchanges to tear one down. A typical Web page, containing an HTML body and numerous embedded images and objects, uses several separate TCP connections. The result is a large RTT multiplier even when there are no network disruptions. The exact value of the multiplier depends on many different factors, such as the size of the page, the number of objects on the page and their size, and browser and server technologies. As many as 30-50 round trips can be considered common with Web-enabled applications.

Another consideration is the extreme sensitivity of application response times to Internet congestion and disruption, which manifests itself as packet loss. The connection points between different networks known as peering points can be a source of packet loss. The networks peer out of necessity but are often in direct competition with their peers which can result in participants' rate limiting incoming bandwidth from another network because of alliances and conflict, resulting in congestion and packet loss. Poor Internet route selection is another source of packet loss and congestion. Any lost or out-of-order packets can result in retransmissions and TCP timeouts, thereby further increasing the RTT multiplier.

"First-mile" bottlenecks, the ability of origin server infrastructure to scale to meet end user demand, such as TCP/IP connection terminations, server load balancing, content compression and SSL encryption/decryption can be addressed by origin server offload techniques. And "last-mile" bottlenecks, the bandwidth capacity of an Internet connection, is increasingly becoming less of an issue as more users adopt high speed broadband connections. This brings "middle-mile" performance bottlenecks to the forefront, namely optimizing both RTT and RTT multiplier, as the key culprit limiting acceptable application response times to a global user base.

Common Approaches to Improving Application Delivery

While companies continue to put more business processes on the Internet through Web applications and services, a number of solutions and technologies have attempted to overcome application delivery challenges.

Potential solutions must be agnostic to content type—dynamic or static, be bi-directional—upload or download, and support not just browser-to-server interactions but Web Services for server-to-server interactions. Furthermore, solutions geared towards improving application delivery must tackle core inefficiencies impacting the RTT and RTT multiplier effect. This requires a set of techniques capable of optimizing at three discrete layers—(1) routing, (2) transport and (3) application—all of which work in combination to provide optimal application response times.

> *Application delivery solutions optimize at three discrete layers—routing, transport and application.*

Approaches to improving application delivery can be categorized as one of the following:

- Data Center Buildout
- Application Delivery Appliances
- Traditional Content Delivery Network (CDN) Services
- Application Delivery Network (ADN) Services

Data Center Buildout

The brick and mortar approach to solving poor performing Web applications is to build out bigger and more data centers thus adding servers and bandwidth to handle demand while moving applications and content closer to geographically dispersed end users. For example, companies may deploy a data center on each U.S. coast, one in EMEA and another in Asia-Pac. There are several problems with this brute force approach. First, building out infrastructure to handle peak demand is expensive and will result, at times, in idle, under-utilized capital assets. In addition, while adding data centers does alleviate middle-mile bottlenecks for those users in close proximity to a data-center, it introduces the need for data replication and synchronization which add cost, complexity and compliance risks. The approach is in stark contrast to the prevalent trend to consolidate servers. In addition, data-center build out is not fundamental to corporate strategy.

Application Delivery Appliances

Specialized appliances have emerged to address inefficiencies in data-center buildout, increase scale and optimize application delivery. Application Delivery Appliances fall into two main categories: Application Delivery Controllers (ADC) and WAN Optimization Controllers (WOC).

ADCs were designed to provide Layer 4 through 7 switching and now provide an array of functions including local load balancing, SSL offload and compression. These devices are single-ended solutions as they reside within the enterprise's data center in front of the web servers.

These devices provide some performance improvement but offer limited value to enterprise-class Web applications which are bi-directional in nature. Bi-directional applications are best served by bi-nodal delivery methods, but it is not plausible to deploy an appliance everywhere a user can access a Web-browser. Because an Application Delivery Controller's footprint is limited to the data center, it cannot offer bi-nodal optimization by placing processing intelligence or services close to end users. This limits their ability to adequately optimize "middle-mile" bottlenecks, resulting in suboptimal RTT and RTT multipliers.

It is not plausible to deploy an appliance everywhere a user can access a Web-browser

WOCs are bi-nodal solutions where appliances or software-based clients reside on both ends of a WAN link to provide shaping, compression and protocol optimization in order to improve application response times. It also helps to improve the performance of branch offices that struggle with certain applications overwhelming WAN links. This is a plausible solution for Intranets with a small number of locations; however, it does not scale well for a large number of locations as IT needs to deploy and manage WOC solutions at every end-user location. It also will not help with extranet applications or Web service queries on the Internet where the end user community resides outside of the corporate WAN environment such as a business partner or customer.

For enterprises deploying Application Delivery Appliances, upfront capital costs are only the beginning of the total cost of ownership (TCO) calculation. Other costs to consider include hardware maintenance, updates, replacement costs, need for additional IT staff, longer deployment time, increased time to value and technical obsolescence.

Traditional Content Delivery Networks (CDNs)

Traditional CDNs cache static content closer to end users, usually with a centralized architecture containing a small number of server locations. Traditional CDNs do not address root middle-mile bottlenecks associated with dynamic Web-enabled enterprise applications and therefore do not help applications like those offered by SAP, Oracle and Outlook Web Access, which have very little, if any, cacheable static content.

Application Delivery Network (ADN) Services

Application Delivery Network Services have emerged as a comprehensive Internet-based platform for improving Web application performance. ADNs have transcended traditional CDNs and Application Delivery Appliances by tackling both first mile and middle-mile bottlenecks and by optimizing delivery of dynamic applications, static content and Web services. By implementing an overlay network to the Internet, ADN service providers use Internet intelligence while combining techniques employed by Application Delivery Appliances and CDNs to transform the Internet into a high performance delivery platform for Web and IP-enabled applications.

ADNs can provide local response times to global users, high availability, on demand scalability and enterprise security with no changes to applications or data-center infrastructure. Guaranteed application performance and availability Service Level Agreements are typically offered; at the same time, TCO is lower than with appliance-based approaches. Costs are more predictable—it's the cost of the monthly service.

The Akamai EdgePlatform—An Application Delivery Network

The Akamai EdgePlatform is a purpose built Application Delivery Network that accelerates delivery of applications and content over the Internet, which can lead to increased usage, adoption and improved productivity. Utilizing a global network of specialized servers connected through optimized protocols, the platform transparently overlays the Internet and optimizes application and content delivery on-demand.

By addressing the shortcomings in the core Internet protocols, Akamai has created a delivery system across the edge of the network that is designed to improve the performance, availability and scale of dynamic IP-enabled applications, Web Services and static content.

A closer look at the Akamai EdgePlatform architecture and technologies reveals why it is the ideal application delivery method for delivering local response times to global users. To better understand the technologies powering the EdgePlatform, it is important to review its key architecture components and how applications are delivered.

Highly Distributed Server Platform

The Platform consists of a bi-nodal overlay network consisting of specialized servers called Akamai "Edge servers". These Edge servers are highly distributed with locations in very close proximity to end users as well as near the origin infrastructure. The platform is a globally distributed computing platform with more than 28,000 servers distributed in approximately 1,000 networks, and approximately 70 countries. With the platform, 85% of the world's Internet users are within a single network hop of an Akamai Edge server.

> *85% of the world's Internet users are within a single network hop of an Akamai Edge server*

The widely distributed nature of the Edge servers implies that for every Internet user and corresponding centralized application, regardless of their location, there is an Edge server region in close proximity. Akamai uses an intelligent dynamic mapping system to direct each end-user and origin location to an optimal Edge server, in essence serving as an on-ramp and off-ramp to the Akamai network. The mapped Akamai Edge servers form a direct bi-nodal network between centralized applications and users across the edge of the Internet. By having servers as close as possible to end users, latency and packet loss are minimized. This is fundamental to optimizing middle-mile bottlenecks, namely RTT and RTT multiplier, as routing and TCP throughput is largely governed by these two parameters.

Bi-Nodal Architecture

When an end user accesses a Web application optimized by Akamai, the browser request is processed by Akamai's dynamic mapping system, which directs the request to a nearby Edge server—a.k.a. "ES-User". By having tens of thousands of globally distributed servers, Akamai's footprint allows for close proximity to end-users, providing the foundation for providing optimal performance as it is critical to get as close to the end-user as possible, while not actually being on the client itself.

Similarly, an optimal origin Edge server in close proximity to customer's application server is also mapped— a.k.a "ES-Origin". The application request goes from ES-User to ES-Origin, which acts as a proxy for the application server. The ES-Origin acknowledges the browser session and the server then issues the request to the application server, obtains the response and forwards the first response back down the line. At this point, two bi-nodal Akamai Edge servers, ES-User and ES-Origin, have transparently interjected themselves into the path between the client and the application server with no change in the transfer process or application infrastructure.

Within this bi-nodal architecture framework, there are several key technologies that are applied to accelerate the delivery, increase availability and improve the scalability of applications.

Akamai SureRoute—Reduces RTT

SureRoute Eliminates BGP Inefficiencies

Akamai SureRoute is designed to remove the inefficiencies of BGP by leveraging Akamai's network of Edge servers and proprietary algorithms to provide a real-time weather-map of the Internet in order to make performance-based routing decisions. At any given time, for each independent user, SureRoute determines a high-performing and available path to communicate between two Akamai Edge servers. SureRoute is beneficial in two ways:

- Optimizes RTT instead of next-hop routing decisions made by BGP, the Internet's core routing protocol. This is increasingly important for those applications where the RTT multiplier is small such as Web service calls, AJAX enabled applications and real-time applications such as VoIP.
- Optimizes application availability of the Internet itself by ensuring end-user requests can reach the application server regardless of Internet bottlenecks such as service provider blackouts, brownouts, de-peering, network outages, earthquakes, etce.

Optimized routing decisions are updated real-time with SureRoute as Internet conditions constantly change. Any communications across the bi-nodal network of two Akamai Edge servers take place over an optimized SureRoute path to ensure optimal RTT for every round-trip taken over the Internet.

Akamai Protocol—Reduces RTT Multiplier Effect

The Akamai Protocol is designed to remove the inefficiencies of TCP and HTTP by leveraging Akamai's network of Edge servers. The Akamai Protocol addresses those middle-mile bottlenecks impacting the RTT multiplier by eliminating the chattiness within the core Internet protocols with the substitution of the more efficient Akamai Protocol as a communication means between origin and end-user Edge servers.

Requests arrive at ES-User as standard HTTP(s)/TCP traffic and are then converted to the Akamai Protocol for travel over the long-haul segments of the Internet to ES-Origin. All Akamai Protocol communications between the two Edge servers take place over an optimized SureRoute path. Traffic is converted back to standard HTTP(s)/TCP once it arrives at ES-Origin.

The Akamai Protocol also benefits by the reduced RTT associated with SureRoute as latency and packet loss govern the maximum data throughput rates associated with the protocol. In essence, the two working in concert provide maximum value than either one could provide by itself.

.

Akamai Protocol Eliminates TCP Inefficiencies:

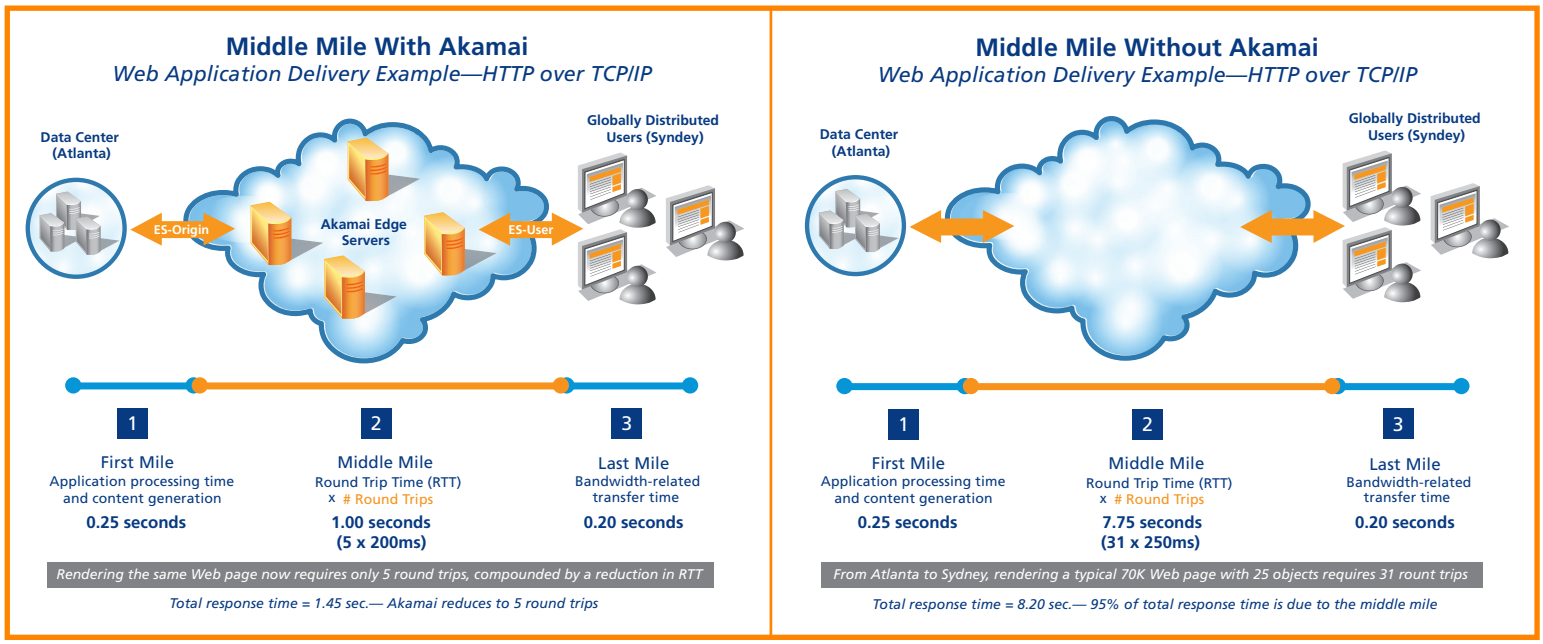
- No three way handshake for connection establishment and teardown—ES-Origin establishes a set of long lived persistent communication connections between itself and Akamai Edge servers. These connections are available on demand for handling multiple browser or machine requests.
- No slow start—Edge servers maintain detailed knowledge of network conditions such as band width and latency. With this awareness, they can communicate immediately at an optimal window size avoiding TCP's slow start mechanism and dynamically change based upon network conditions.
- Guaranteed pipelining—leveraging the persistent connections and intelligence between Edge servers and the origin server's data center, Akamai employs pipelining that allows multiple HTTP requests to be multiplexed over a single connection without waiting for a response.
- Intelligent retransmission—Edge servers maintain a significant amount of information such as latency between machines, transmission window sizes and packet sequencing information and thus provide a more intelligent retransmit methodology than the TCP timeout parameter.

TCP has an aptly named "slow start" feature

Akamai Protocol Eliminates HTTP Inefficiencies

HTTP, the browser's core protocol, magnifies TCP inefficiencies by requiring multiple TCP connections be established and torn down to deliver a page.

- Intelligent pre-fetching—when ES-User delivers the base page request to the browser, it simultaneously parses recursively the HTML base page along and predicts requests for subsequent embedded objects and immediately issues the corresponding requests to the origin server. All of the content is then transferred back as a single transaction to ES-User using the proprietary Akamai Protocol. When the browser receives the base page and then requests the remaining elements of the page, they are already waiting at the ES-User and are delivered as if the origin server were only a few milliseconds away.
- Compression—data is compressed en route, reducing bandwidth usage.
- Caching—any cacheable content is stored at the ES-User close to end users and served from cache.



Origin Infrastructure Offload

Akamai Edge servers provide several "first-mile" performance optimization techniques to offload process intensive application server tasks to better ensure application performance and scale:

- TCP Connection Management—by off-loading the client TCP/IP connections to ES-Origin, the origin server processing resources can be devoted to serving content. Origin server performance can also be enhanced by creating multiple persistent TCP/IP connections between ES-Origin and the origin application server.
- SSL Termination and Offload —customers can choose to have their origin servers transmit in the clear and have Edge servers secure transmissions through SSL to the browser. Through this configuration, customers get the benefits of SSL while offloading the SSL processing to Akamai's infrastructure — increasing origin server scalability and performance.
- Compression—Edge servers dynamically compress and decompress data that will benefit from compression while reducing the payload needed to be transferred across the Internet. Text objects such as HTML and Java script are highly compressible—often able to be compressed by 90%.
- Caching—Edge servers can be configured to automatically cache static content, such as small page objects (e.g., small .gifs, .docs, and .jpg), documents or digital assets such as software or media files. This enables the Akamai Platform to offload the origin infrastructure and reduce page response time by serving the cacheable content in close proximity to the end-user. Akamai honors industry standard cache lifetime controls and heuristics.

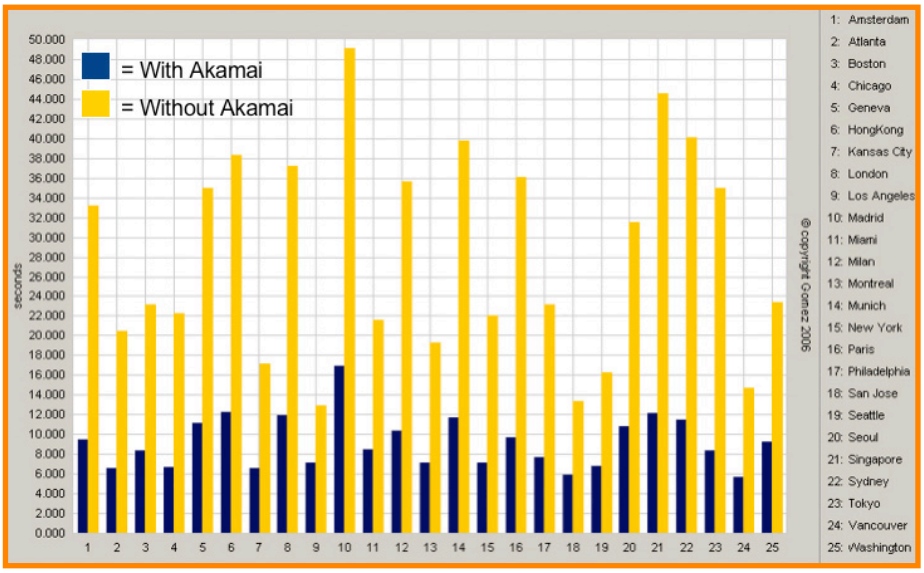
Proven Application Performance and Scalability

Akamai's highly distributed EdgePlatform, real-time SureRoute route prediction and the high-performance communication Akamai Protocol work in concert as a system to improve the performance and scalability of dynamic applications and Web Services. By addressing the Internet's root protocol problems, dramatic performance gains can be achieved regardless of application type or class.

Akamai Helps to "Flatten the World"

Users feel as if they are close to a data center, regardless of location

Example: Customer Support Portal—4 Step Dynamic Transaction



Business Impact

The Akamai EdgePlatform is a powerful extension to any enterprise Web infrastructure. With the ability to optimize delivery of dynamic applications and Web Services, enterprises can centralize their infrastructure without compromising the performance and availability of their application whether their delivery requirements are coast to coast or global. No changes to Web applications, servers or clients translate to low risk and fast time to value.

The Akamai Platform delivers on five critical requirements in a single comprehensive platform for all classes of Web applications:

- LAN-like performance for global users—globally distributed Edge servers, SureRoute and the Akamai Protocol work together to enable LAN-like response times to global users for dynamic applications and Web Services.
- High availability—- dynamic content that must be retrieved from the origin is protected by SureRoute's ability to avoid Internet bottlenecks in real-time.
- Predictable capacity—meets the demand for dynamic applications and Web Services that can spike to thousands of times normal traffic levels at any given time.
- Enterprise Security—provides secure and accelerated delivery of SSL applications while integrating seamlessly with existing authentication systems and process such as the support for X.509 client based certificates and PCI compliance for eCommerce.
- Minimal Total Cost of Ownership (TCO)—no hidden costs as TCO equals the cost of the monthly subscription. Zero application and infrastructure modifications are required, translating to fast time to value and low risk.

In a recent White Paper published by IDC—"Determining the Return on Investment of Web Application Acceleration Managed Services"—the average annual benefit experienced by an organization using Akamai's EdgePlatform for Web-based enterprise application delivery was $42,830 per 100 application users. Based only on the cost reduction benefits, the payback period from deploying the service averaged 1.8 months for the companies surveyed, yielding an average return on investment of 582%.

Summary

The Akamai EdgePlatform addresses the root-causes of Internet performance and scale bottlenecks by intelligently placing itself between the end-user and origin data-center as an overlay network to the Internet, transforming it into a high-performing and reliable application delivery platform. Along with delivering one-of-a-kind protocol acceleration techniques such as SureRoute and the Akamai Protocol to address middle-mile performance bottlenecks, Akamai's Application Delivery Network delivers the benefits of Application Delivery Controllers, WAN Optimization Controllers and Content Delivery Networks in a single platform that elegantly combines multiple technologies to deliver on-demand performance and scale to dynamic applications and Web services.

While an increasing number of enterprise applications are becoming Web-enabled, there are other applications such as legacy traditional client/server, virtualized and real-time sensitive applications not well suited for a Web interface. Such applications are also vulnerable to middle-mile bottlenecks in an analogous manner. The use of Akamai's distributed EdgePlatform, SureRoute capabilities and Akamai Protocol can be extended to improve any IP-enabled application whether Web-based or not to leverage the economics, scale and reach associated with the Internet as an application delivery platform.

About Akamai

Akamai® is the leading global service provider for accelerating content and business processes online. Thousands of organizations have formed trusted relationships with Akamai, improving their revenue and reducing costs by maximizing the performance of their online businesses. Leveraging the Akamai Edge Network, these organizations gain business advantage today, and have the foundation for the emerging Internet solutions of tomorrow. Akamai is "The Trusted Choice for Online Business."
For more information, visit www.akamai.com.

AKAMWP-ACC-APP-1207

Akamai Technologies, Inc.
U.S. Headquarters
8 Cambridge Center, Cambridge, MA 02142
Tel 617.444.3000
Fax 617.444.3001
U.S. toll-free 877.4AKAMAI
(877.425.2624)

Akamai Technologies GmbH
Park Village, Betastrasse 10 b
D-85774 Unterföhring, Germany
Tel +49 89 94006.0

www.akamai.com