



Content Delivery and the Mirror Image Adaptive CAP Network

A Technology White Paper

| | |
|--|-----------|
| Introduction | 3 |
| Performance Results..... | 3 |
| Infrastructure Investment | 4 |
| Content Delivery Approaches..... | 5 |
| Caching Architecture..... | 5 |
| Dispersed “Convenience Store” Architecture | 5 |
| Dedicated Architecture..... | 5 |
| Concentrated “Superstore” Architecture | 6 |
| The Optimal Solution – Mirror Image® Internet | 7 |
| Content Access Point® (CAP) Network..... | 7 |
| CAP Connectivity..... | 9 |
| CAP Architecture | 9 |
| How the CAP Network Works..... | 9 |
| CAP Benefits | 10 |
| Advanced Content Delivery, Streaming Media and Web Computing Solutions | 11 |
| Conclusion..... | 11 |

Introduction

Speed at the desktop has been solved by Moore's Law, which gives us continuously faster processors and cheaper memory. Laying endless miles of fiber-optic cables and sending data packets via high-speed, fiber-optic transmission technologies such as Dense Wave Division Multiplexing (DWDM) have solved bandwidth in the network. However, as Internet traffic levels climb inexorably upward and bandwidth-intensive applications proliferate, the issue of boosting Internet content delivery and solving the Internet latency, resulting from overloaded or unresponsive origin servers, traffic-clogged routers, packet loss, bottlenecks and outages along the Internet backbones, cannot simply be solved by increasing bandwidth or end-user processor speed.

Granted, most Internet Service Providers (ISPs) offer excellent network operations, but they cannot control the entire Internet. And, while a business may choose a great company to host their Web site, that company is also not able to control the whole Internet. After all, the Internet is an extremely complex roadway comprised of thousands of communication companies from diverse backgrounds, including telephone, data, cable and energy. This complicated mapping of connections is constantly being upgraded. At any one moment, thousands of changes are being made around the world.

Today, many Web content delivery service providers offer solutions that attempt to boost content delivery and resolve Internet latency. Some have deployed dispersed networks with thousands of locations placed very close to end users; others have deployed a more concentrated Internet architecture that has been strategically situated between end users and the Internet backbone worldwide.

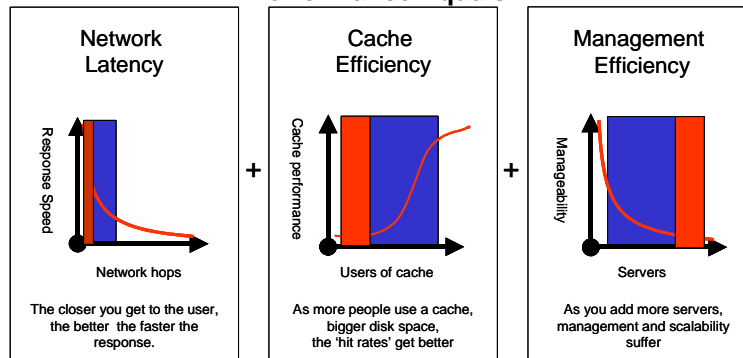
In a Gilder Technology Report, George Gilder uses the phrase "7-Elevens" of the content distribution space to describe dispersed networks with many small locations, "...sure it's convenient to stop by and pick up the morning paper and a Big Gulp on the way to work, but notice the manager's face when you ask him for directions to the shoe department." In this "convenience store" model, the cache servers, constrained by space limitations, are small. As a result, they have limited storage so they must prioritize cache content. Often, as traffic increases, the routes they map across the Internet to find user requested content become more convoluted because the average data packet must contend with many more routers, network hops and data traffic jams.

With fewer and larger enterprise-class servers, the concentrated architecture, as Gilder likes to refer to as the "superstore" model, differs from the dispersed "convenience store" model because it is not forced to prioritize content or limit cache size. As a result, it provides more processing power and storage than the smaller footprint of the dispersed store.

Performance Results

When it comes to content delivery, the ability to improve performance depends on several variables, including network latency, cache efficiency and management efficiency.

Performance Equals:



Network latency is attributable to Internet traffic, as well as the number of Internet nodes the content must cross to reach its destination. Locating the content closer to the end user can reduce latency however it has implications on cache efficiency and management efficiency.

Conceptually, you could replicate your entire Web site near every user, but this is not a practical solution. Content delivery services distribute content close to users by storing commonly requested objects in cache (an

object is in cache when it has been previously requested by a user). Performance is improved if the object is in cache when a user requests it. For example, some “convenience store” architectures may have as many as 500 times more servers as a “superstore” infrastructure. Therefore, in the “superstore” model each server would receive 500 times more traffic than a server in a “convenience store” model serving the same visitors. The probability that a particular page would have already been requested is much higher in the “superstore” infrastructure, so any particular object is more likely in cache when it is requested, avoiding additional routing and resulting hops.

The more concentrated “superstore” model also adds greater management efficiency. Configuration or content changes can be deployed faster. For example, when an object is modified, it would take much longer for the change to be applied across the vast, intricate set of servers in the “convenience store” network. As a result, it’s more likely that your visitors would view outdated content.

Infrastructure Investment

With the smaller cache size, the “convenience store” model also requires more origin server requests. For example, the “convenience store” network with 10,000 servers would require 50,000 origin server requests for the initial cache load of a page with 5 objects, compared to 100 requests from a “superstore” model with 20 servers.

Due to its smaller size, the “convenience store” cache can often only maintain each object in cache up to 6 hours, compared with up to 6 days in the “superstore” cache (i.e., cache content must be flushed to manage limited storage capacity, particularly with larger objects). In a 6-day period, the “superstore” cache would reload cache 24 times (4 times each day for 6 days). The “superstore” server could maintain the same objects for the entire period. Over the period of a month (5 x 6 days), the origin server would receive monthly 6 million requests from the “convenience store” network. In contrast, the “superstore” network would generate 12,000 requests during the same period (99% fewer origin server requests). This translates into a smaller processing burden for the origin server and significant savings in the Web site infrastructure investment required to support the traffic.

Cache Content Requests

| | “Convenience Store” Model (10,000 servers) | | | | “Superstore” Model (20 servers) | | | |
|------------------------------|---|-----|---|-----------|------------------------------------|-----|---|--------|
| | Content Retrieval | | | | Content Retrieval | | | |
| Initial Cache Load | | | | | | | | |
| Number of caches | 10,000 | | | | 20 | | | |
| Objects per page | X | 10 | = | 50,000 | X | 10 | = | 100 |
| Monthly Cache Refresh | | | | | | | | |
| Number of caches | 10,000 | | | | 20 | | | |
| Objects per page | X | 10 | = | | X | 10 | = | |
| Monthly reload frequency | X | 120 | = | 6,000,000 | X | 120 | = | 12,000 |

After discussing various content delivery architectures, this paper focuses on the concentrated “superstore” approach to boosting Internet performance and discusses how Mirror Image® Internet builds upon this architecture with an intelligent, cost-effective solution that optimizes online content, transaction and application delivery.

Content Delivery Approaches

Four differing models attempt to optimize Internet content delivery.

In today's marketplace, four different network approaches – caching, dispersed, dedicated and concentrated architectures – attempt to solve Internet content delivery. An overview of each is provided below.

Caching Architecture

The caching architecture uses pre-positioning in combination with caching to place medium sized content servers in local (tier 3 or 4) or regional Internet Service Provider (ISP) Points of Presence (POPs). Each of these servers connects back to a central controlling server. This architecture handles a limited number of users, serves content specific to each ISP and places tremendous demand on bandwidth because user requests may be served from the local cache, central content repository or the origin server. In addition, as a result of scalability problems, companies that have deployed this model tend to focus on very specific content delivery market segments, such as enterprise networks and streaming. Also, because this architecture is heavily dependent on bandwidth, most Tier 1 carriers that derive a significant portion of their revenue from bandwidth sales have either deployed the caching architecture through an in-house development effort or through a third-party content distribution tools provider.

Dispersed “Convenience Store” Architecture

The dispersed “convenience store” model distributes small content servers at POPs across local ISP networks. When a server receives a request for content from an end user, it refers to an internal mapping table to locate the content, which may reside at a POP within the same or on another ISP network. While this architecture offers the advantage of placing servers close to end users, servers often have to contend with Internet exchange (IX) point congestion, as well as the Internet latency that may result from requested content being retrieved from a POP in another region or from the origin server.

Dedicated Architecture

The dedicated model places small content servers across regional ISP networks that are connected by a dedicated network. Although this model does not encounter significant delays due to Internet congestion, it does require significant server processing time and power to coordinate the transmission of data between regional networks.

Each of three aforementioned architectures share the following disadvantages:

Finite disk space—With a finite amount of server disk space, content is typically dispersed and stored among multiple servers across a numerous POPs. As a result, it takes significant processing time and power to determine which server is storing user requested content. This situation may delay content delivery and slow response time by creating excessive network overhead.

Difficult to scale—Due to the limited disk space and capacity, especially in the dispersed architecture, servers often need to be upgraded to handle content needs. As a result, the caching, dispersed and dedicated architecture models may prove extremely difficult and expensive to scale.

Networking and management complexity—When trying to improve Internet performance, complexity becomes inherent for each of these models. For example, if a server runs out of disk space, it may be forced to compensate by taking one of two actions:

- 1) Try to extend the size of the server by combining thousands of smaller servers so that it appears like a larger, central machine.
- 2) Discard a current document and either connect to another networked device or the origin server in the hopes of immediately filling a user request or serve the content the next time a user requests it (after responsibility for serving the current request is “passed” to another server).

Conflict with networking design principles—From a networking design perspective, the caching, dispersed and dedicated models require each server have a sufficient number of users at each level in the network, so a fairly wide, flat hierarchy of servers often proves to be the most appropriate design. However, this type of network

design recreates, for content delivery, the mistakes that were implemented during the evolution of Internet networking architectures. In fact, it directly conflicts with the optimal “best-in-class” hierarchical network design principles in practice today to simplify routing, where backbone routers are at the top of the hierarchy and connect into smaller routing devices or layer 3 switch/routers. From there, these devices connect into large carrier class layer 2 switches that use Virtual Local Area Network (VLAN) technology to create different networking segments. Each VLAN port sets up a separate physical or virtual network to create not only a hierarchy at the physical level but also a hierarchy at the IP networking level. These switch devices then connect into smaller switches that feed into smaller hubs that connect to individual user workstations or devices.

The aforementioned level of hierarchy was developed as a result of years of trial and error that date back to the 1980’s with the advent of the bridge, the first device that attempted to extend the network by connecting disjointed network portions to create a vast and fairly wide, flat hierarchy. This approach quickly proved inadequate for solving the growth and scale issues of corporate and ISP networks. So why should it be any different for the caching, dispersed and dedicated architectures described above? All of these models will not be able to adequately scale to handle the Internet’s expected growth levels, especially as bandwidth intensive applications continue to proliferate.

Use “The Problem” to solve “The Problem”—Perhaps the greatest disadvantage for the caching, dispersed and dedicated architectures is that they use “the network” and network connectivity as the main element to solving the problem. However, the problem concerns network issues such as traffic-clogged routers, packet loss and outages along the Internet backbones; bottlenecks that develop during busy periods; and “speed-of-light” latency. However, the caching architecture uses “the network” to pass messaging back and forth; the dispersed architecture uses “the network” to pass content from one local POP machine to another; and the dedicated network architecture uses “the network” to pass content between machines located at regional POPs. Is it really feasible to use “the network” to solve a problem that is “the network”? Isn’t there something fundamentally flawed in a solution that uses the problem to solve the problem? These three architectures attempt to solve the problem for the end user by bringing back the complexities of the network.

As traffic levels on the Internet continue to climb and the use of bandwidth intensive applications expands, the aforementioned caching, dispersed and dedicated network architectures will fall victim to the issues and design flaws presented above. Ideally, a more optimal architecture for Internet content delivery would leverage the advantages of maintaining content in closer proximity to end users by having an adequate amount of disk and processing capacity to store and deliver every piece of content that is requested more than once. Such architecture would never retrieve a second copy of an unchanged document and therefore, would generate the least amount of network traffic. The fourth concentrated “superstore” architecture, detailed below, is most in line with this goal.

Concentrated “Superstore” Architecture

In the concentrated “superstore” model, Internet content is placed in massive servers in central locations close to large regions of users, specifically, at IX points. Content requests are stored in cache at each central location, thereby avoiding the congestion of links within North America and the expense of international links outside North America. Furthermore, the user base around each concentrated grouping of servers is large enough to ensure a higher number of duplicate content requests. In a concentrated architecture environment, ISPs and other corporate organizations are able to easily connect to the central storage location by simply adding an additional higher layer to their current hierarchical network design. As a result, organizations are able to maximize the benefits of content delivery by serving users from these massive central storage locations.

While a significant improvement over the caching, dispersed and dedicated architecture models, a concentrated architecture does not always solve the performance and cost efficiency issues associated with Internet content delivery, such as transaction based processing. However, it does provide a stronger, more intelligent foundation for content delivery in the Internet’s ever-changing landscape.

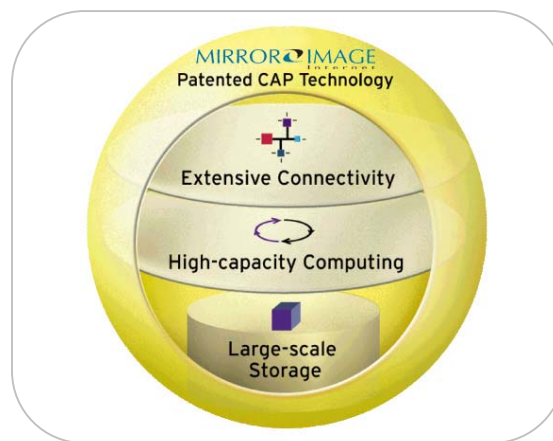
The Optimal Solution – Mirror Image® Internet

Scalable, concentrated platform combines optimal mix of connectivity, processing power and storage to power advanced Content Delivery, Streaming Media and Web Computing services.

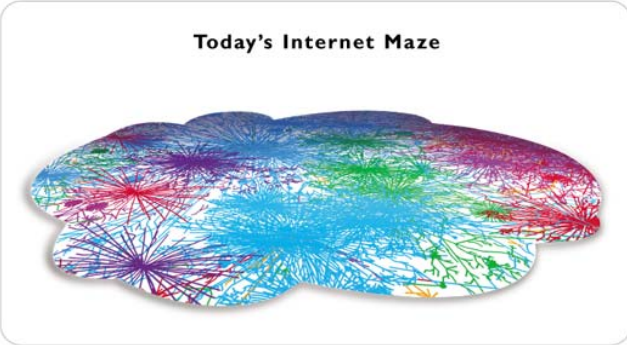
Based on the concentrated architecture detailed above, Mirror Image® Internet, a leading provider of advanced Content Delivery, Streaming Media and Web Computing solutions, has deployed an adaptive Content Access Point® (CAP) network that provides reliable content, application and transaction delivery to users around the world.

Content Access Point® (CAP) Network

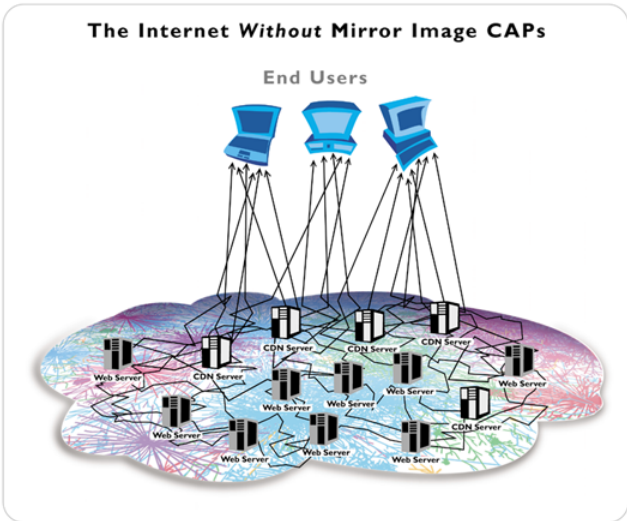
Unlike other architectures that rely on highly dispersed networks of small server appliances, the modular design of Mirror Image's global CAP network combines an optimal mix of connectivity, processing power and storage to provide a scalable, concentrated platform for content, application and transaction delivery. Comprised of technology developed and patented by Mirror Image, accompanied by best-of-breed equipment from leading network vendors, the global CAP network offers the scalability, security, redundancy, control and CPU capacity expected of an enterprise-class system. The resulting platform provides secure and integrated content distribution, proxy, streaming, storage, download, verification, monitoring, reporting and tracking capabilities as well as access control, application distribution, transaction processing, logging and billing support.



Robust and reliable, the CAP architecture is built from Mirror Image patented technology operating on state-of-the-art equipment from leading vendors.



The Internet is a hectic place, crowded with a growing population of users, obstructed by overburdened servers and clogged routers, and bogged down by bandwidth-hungry applications.



When an end user visits a Web site, the quest for content is not always easy. Before content can be delivered, the visitor's request needs to travel through the Internet's maze of connection points, where it's often at the mercy of the weakest link.



Mirror Image's global CAP network rises above the Internet to provide a scalable, concentrated platform that reliably delivers Web content and applications to users around the world. As a secure and managed high-speed layer on top of the Internet, the CAP network offloads origin servers and infrastructure by intelligently placing content and applications closer to millions of users worldwide. This concentrated approach bypasses Internet congestion to improve site performance, enabling organizations to provide users with a seamless Internet experience 24 hours a day, 7 days a week.

CAP Connectivity

After evaluating demographics, Internet usage and connectivity issues in numerous regions around the world, Mirror Image deployed the CAP network at strategic IX peering points that had the highest concentration of end users, Web traffic and network convergence. As a secure and managed high-speed layer on top of the Internet, the CAP network reduces Internet hops, bypasses congestion and circumvents overloaded peering points to provide a faster, more reliable solution to users, regardless of location or demand.

CAP Architecture

The modular design of the CAP architecture combines internally developed and patented Mirror Image technology with enterprise-class equipment from leading vendors, including Cisco, Hewlett-Packard, Oracle and Sun. Coupled with extensive connectivity, the concentrated deployment of the CAP platform ensures a consistent, predictable and stable level of service that quickly and securely leverages the following to deliver fresh content with unmatched performance, scalability and reliability:

Internet Gateway

The CAP network combines local content delivery with high-speed direct cross connects, a physical piece of wire or fiber optic cable that creates a high-speed connection between two networks and terminates at a router or Internet gateway, to boost Internet performance and improve the efficiency of connected networks.

Core Components

The core of the CAP architecture serves as the central point for all user requests for content and is able to reliably handle thousands of requests in record speed. The CAP core is comprised of:

Mirror Image Intelligence: Active Content Verification (ACV) algorithm, designed and patented by Mirror Image, ensures the reliable delivery of the freshest content available, as well as provides detailed traffic analysis statistics and tracking to guarantee content freshness on a massive scale.

High-capacity computing: A sophisticated combination of enterprise-class Hewlett-Packard servers running Oracle databases, coupled with Gigabit Ethernet connectivity that uses high-speed fiber channel connected disk arrays.

Massive terabyte-sized disk storage: High-speed fiber channel disk arrays provide massive capacity for storing Internet traffic and Web content.

Service Components

The following dedicated infrastructure and enterprise servers provide a front-end to the CAP core and power Mirror Image's advanced Content Delivery, Streaming Media and Web Computing services:

Feeder Subsystems: The feeder subsystems connect the CAP object storage to the customers network via the reverse proxy subsystem. Acting as middle-tier caches to the Oracle databases in the core, the feeders move content from the Oracle databases to the reverse proxy subsystem.

Getter Subsystems: The getter subsystems focus on pulling content to the Core database, minimizing the number of content requests that must be routed to the origin server.

Reverse Proxy Subsystem: The reverse proxy subsystem provides the primary connection between the CAP network and the customers network. Being a cache itself, this HTTP proxy is able to serve the content from its own cache or if needed requests content from the feeder subsystem.

Load Balancing: All subsystems are load balanced at the switch, which provides redundancy and scalability to the CAP architecture.

How the CAP Network Works

When a user requests content from a Mirror Image provisioned Web site, the distributed URL automatically routes the request to a global load balancer on the CAP network, where DNS racing determines the IP address of the CAP location that can provide the fastest response time. Once the selected CAP location receives the request, the caches, and then the core database, are checked for the requested content. If the content is found, it is automatically delivered to the user, thereby completing the transaction. If the content is not found, the CAP network automatically returns a redirection status code of 302 to the origin server's URL and the requested

content is automatically delivered to the user from the origin server. At the same time, the CAP network retrieves, or “pulls”, the requested content from the origin server and stores it, making it available for all subsequent user requests. This method represents Mirror Image’s “Pull” model, where objects are “pulled” from origin sites and stored on the CAP network after the first user request. This method eliminates the processing power required to distribute content that might never be requested by users. In addition, in certain cases, customers can opt to “push” content to the CAP network. This method ensures that even first user requests will be served directly from the CAP network, eliminating the need for the customer to store the objects locally.

CAP Benefits

Powering Mirror Image’s advanced Content Delivery, Streaming Media and Web Computing solutions, the adaptive CAP network intelligently places content and applications closer to millions of users worldwide. By aggregating and storing content and providing an outsourced platform for content, application and transaction delivery, the CAP network cost-effectively manages traffic and efficiently scales to reliably serve user requests. The end result improves site performance by offloading origin servers and infrastructure, quickly delivering online content and applications, providing transaction support and protecting against site outages and surges in demand. Furthermore, this innovative approach increases resource efficiency and significantly reduces bandwidth, capital and management costs.

Scalable and Powerful

Each CAP location provides a powerful, fault-tolerant platform that offers virtually unlimited scalability and around-the-clock surge control protection in case of sudden, unexpected peaks in traffic. Equipped with hundreds of processors and terabytes of storage, the CAP network’s concentrated architecture easily expands to accommodate user demand and increasingly complex Web content. This capability offers a major advantage, particularly as content, application and transaction delivery proliferates on the Internet.

Easy to Manage

In comparison to architectures that deploy hundreds or thousands of servers, the CAP platform’s concentrated design makes upgrade efforts easier to manage. In fact, built inside a configuration file with one master install package, all it takes is a single install procedure to seamlessly add additional devices to the CAP architecture.

High Availability and Built-in Redundancy

The CAP network delivers fast, non-stop operation to organizations worldwide. Dual storage areas, as well as redundant feeders, getters and reverse proxy servers, maximize performance and availability to provide automatic fail-over. These features, along with a high-speed internal backbone, enable Mirror Image to ensure continuous data delivery and more than enough throughput capacity to handle an unlimited number of requests. Furthermore, separating service components from the CAP core guarantees that each service will continue to function even during the most catastrophic failures. For instance, in the hypothetical and highly unlikely case of a specific CAP location losing its core database, Mirror Image’s DNS services would automatically and transparently resolve the request to a more optimal, and most likely next closest CAP location.

Secure

Mirror Image uses a comprehensive, layered approach for increased security to guarantee the highest levels of protection for all content stored and delivered from the CAP network. With 24x7 monitoring services, Mirror Image proactively ensures the highest levels of security and availability. In fact, Mirror Image engineers set up and manage an array of security-related solutions to recognize and correct attempted attacks and security breaches before they occur. In addition, Mirror Image provides:

- 24 x 7 intrusion detection monitoring and management
- Monthly incident status and audit report
- Dedicated software for intrusion detection
- Prompt notification in the event of system problem occurrences
- Installation of the latest security patches
- 128-bit encryption on all content uploads
- Carefully guarded privilege access control to system equipment and files
- TripWire, an audit trail tool, for detecting intrusions
- TCP Wrappers that close down port access, making computers less vulnerable to attacks
- Tightened security for all available Mirror Image system ports
- The installation of security packages such as TACACS+ and Radius (for controlled access)

Advanced Content Delivery, Streaming Media and Web Computing Solutions

Mirror Image's CAP network powers advanced Content Delivery, Streaming Media and Web Computing solutions that provide accelerated performance, around-the-clock availability, unlimited capacity and flexible control required to maximize revenue opportunities, reduce infrastructure costs and optimize customer satisfaction. And, backed by quantifiable metrics and world-class customer support, the following solutions help organizations maintain a competitive advantage in today's increasingly volatile market:

Content Delivery

Global Content Caching: Offloads processing and protect Web sites from traffic spikes by cost-effectively serving static content, Java applets and content generated by costly backend servers.

Digital Asset Download: Reduces origin site storage, processing and bandwidth costs by providing guaranteed capacity and availability to ensure the fast, secure and reliable download of digital assets (e.g., software, documents and high-definition images).

Business Continuity: Ensures around-the-clock Web site availability by continuously monitoring and automatically serving content in the event that a site becomes unavailable.

Streaming Media

Video On-Demand: A single-source solution for the efficient and reliable streaming delivery of digital content.

Webcasting: Achieves communication goals by allowing users to deliver "one-to-many" messages for training, marketing and distance learning outlets.

Web Computing

Managed Content Targeting: Empowers organizations to globally deliver targeted Web content and marketing campaigns based on specific logic (e.g. geography, language, time of day) without complex programming or Web site changes.

Extensible Rules Engine (XRE): Gives customers control over the delivery process by providing the ability to code logic for execution on Mirror Image's network.

Web Beacon: Enables businesses to cost-effectively evaluate marketing campaigns by anonymously tracking click-stream activity to identify user behavior and trends.

Java Execution Environment: Provides a Java container that enables an organization to distribute Web application logic directly to Mirror Image's network without platform-specific consideration, special syntax or custom programming.

Conclusion

Mirror Image's concentrated CAP architecture provides a leading solution for advanced Content Delivery, Streaming Media and Web Computing.

Although a number of content delivery choices exist, most traditional architectures use complex networking methods to try to resolve Internet latency. Undoubtedly, these models will not be able to reliably handle Web content and Internet traffic growth. Therefore, Mirror Image's use of a concentrated CAP architecture provides advanced Content Delivery, Streaming Media and Web Computing solutions that provide customers with the accelerated performance, around-the-clock availability, unlimited capacity and flexible control required to improve revenue opportunities, reduce costs and increase customer satisfaction.