# How to Cost-Effectively Scale Web Servers

The exponentially growing demand for Internet-based marketing, sales, support, and software distribution services is increasingly driving companies to expand the capacities of their World Wide Web (WWW) servers. While the conventional solution to resolving capacity requirements has been simply to replace an existing server with a bigger and faster model, there is a far more cost-effective and beneficial approach.

That solution includes two products—Cisco's LocalDirector and DistributedDirector. Each of these systems works to protect existing server investments, improve response time for end users, and automatically provide an alternate-path routing capability that is just not possible with the single-server approach. The result: an unprecedented ability to effectively and economically scale Web services.

With the Cisco LocalDirector and DistributedDirector, new servers can be added and removed transparently to end users, because a population of multiple servers can be made to appear as a single virtual server. As a result, only a single IP address—and a single Universal Resource Locator (URL)—may be needed for the entire server complex.

The Cisco LocalDirector tracks local network sessions and server load conditions, directing each session to the most appropriate server at the time. The servers appear as one virtual server with a single IP address.
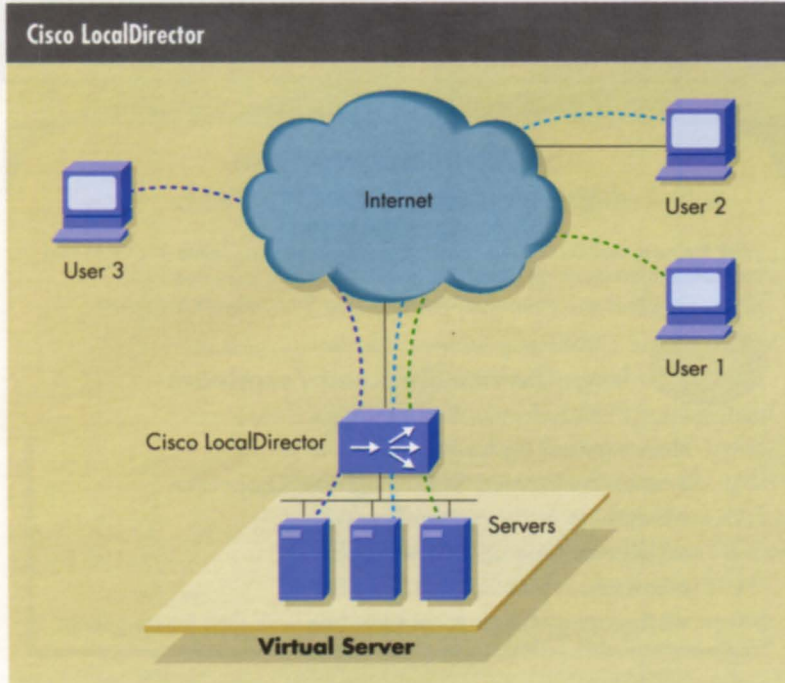
For service providers, this capability offers a way to increase Web server capacity in small, low-cost increments. At the same time, incremental server expansion allows service providers and others to add or remove servers from the network without impacting end users. Equally important, the network manager is empowered to provide end users the timely service they expect by matching capacity with demands and always routing requests to the nearest available servers in order to reduce the number of hops—and potential time delays.

The Cisco LocalDirector, now shipping, transparently and intelligently redirects sessions—as they are established at a Web site to a single virtual IP address—to other local servers as necessary. This redirection is accomplished, independently of both the Domain Name Service (DNS) and the application it serves, by rewriting the IP header information in accordance with a dynamic table of mappings between each session and the server to which it has been redirected. In other words, LocalDirector treats a cluster of servers as one virtual IP address; it is this virtual IP address that DNS registers. LocalDirector can effectively deal with client applications that cache DNS entries for extended periods of time.

When establishing a redirection, the Cisco LocalDirector considers the status of all servers and their ability to receive additional sessions on the basis of their existing loads and availability; servers might be marked unavailable if, for example, they are taken out of service for maintenance. The Cisco LocalDirector performs this analysis without incurring any network overhead and with no specialized host or client software.

Unlike the Cisco LocalDirector, the Cisco DistributedDirector—which will be available in the fourth quarter of 1996—can act either as the primary DNS name-server for a specific subdomain or as a Hypertext Transfer Protocol (HTTP) session redirector for a specific URL. In the former mode, the DistributedDirector sorts responses to name service queries by the relative locations of the user and the server.

Cisco LocalDirector

Internet

User 3

User 2

User 1

Cisco LocalDirector

Servers

Virtual Server

In the DistributedDirector illustration on page 8, if a DNS query for a named service, such as www.cisco.com, is sent from a user in Los Angeles, California, the DistributedDirector would respond with the IP address of the Web site in San Jose, California. But if the query came from a user in Sweden, that user would be directed to the Cisco Web site in Brussels, Belgium.

Operating in "HTTP redirect" mode, the Cisco DistributedDirector masquerades as the Web server defined by the user-specified URL, accepts HTTP connections, determines the best server based on the relative locations of the client and server, and redirects the client to that server.

By directing users to the server nearest to them, DistributedDirector can cut response time dramatically as it minimizes the number of hops. In the near future, DistributedDirector's configurable metrics will be enhanced with an ability to factor in servers' load conditions in addition to geographic proximity.

These parameters can be important performance issues as companies expand their Web presence and the capacity of their Web server platforms. With this growth and the increasing volumes of information and services supported by these platforms, consistency of service, availability, and ease of use are all critical issues that must be addressed. Not to be overlooked is the operational imperative of being able to maintain this high level of service throughout the maintenance process—whether that maintenance procedure calls for adding servers or removing them for repair.

In short, what companies need is a Web strategy that provides a seamless and cost-effective growth path to tomorrow's burgeoning requirements. Cisco LocalDirector and DistributedDirector products provide that growth path, offering a simple, direct way to protect existing server investments while incrementally scaling server capacities and global connectivity options in such a way that site access and customer service are always optimized.

Using the Web for Business?

Cisco Systems' suite of online services and information is growing rapidly and gaining acclaim from the Cisco user community. For information on Cisco's new Internetworking Product Center and other Web-based tools and resources, see "Open for Business!...," page 30.
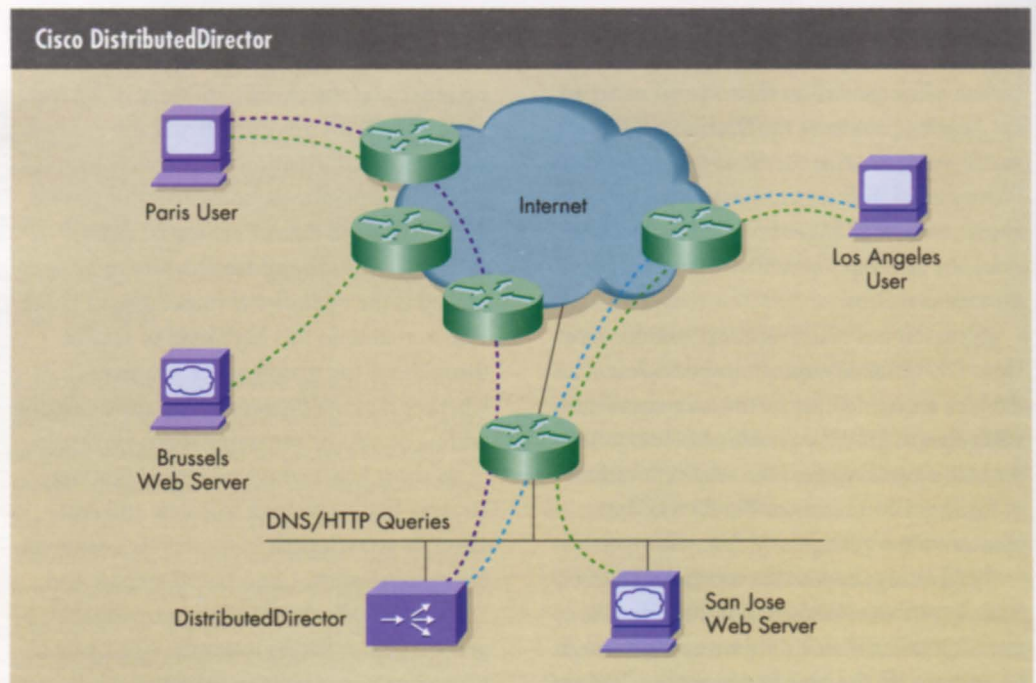
re: How the Cisco LocalDirector Works

To provide TCP connectivity across multiple servers, Cisco LocalDirector relies on an inverse multiplexing mode (IMM) that presents the appearance of a single server to the outside, while actually using the power of multiple servers. The system supports multiple IMM groups, each consisting of a virtual address multiplexed to many actual servers, and each of these servers having a real IP address.

LocalDirector measures the load on the servers specified in the IMM groups and intelligently distributes the load among them using the session distribution algorithm (SDA) developed by Cisco. SDA detects the long response time when a server fails, and redirects pending connections to the most responsive servers. The algorithm then continues to periodically check the

failed server, and when it returns, SDA begins to issue connections again.

LocalDirector also operates in a forward multiplexing mode (FMM) to eliminate the multiple-domains-from-one-server problem by using multiple TCP ports to allocate multiple virtual addresses mapped to a single server. With FMM, a single server can run multiple Web server processes, each residing on a different TCP port, even while the outside world sees the virtual addresses for three separate servers. LocalDirector also allows administrators to scale this configuration by specifying multiple sets of FMM groups. In addition, administrators using this system can specify the physical addresses in groups to take advantage of load balancing, as described for IMM.

Web Scalability, *Continued*

Cisco DistributedDirector

DistributedDirector assesses the relative topological locations of end users and geographically distributed servers to determine the best server for each session, acting as a primary DNS name-server or as an HTTP session redirector.

re: How the Cisco DistributedDirector Works

In operation, the Cisco DistributedDirector can be configured on a per-domain basis either as a DNS caching name-server or an HTTP session redirector.

DNS Caching Name-Server Mode

DNS caching name-server mode can be used for all IP services. In this mode, the Cisco DistributedDirector acts as the primary DNS caching name-server for a specific subdomain. Using the Director Response Protocol (DRP), the DistributedDirector sorts DNS query responses by the information contained in the router-based networking infrastructure. This information describes the relative topological locations of the user and the server.

For example, when a client sends out a DNS query for a named service, such as www.cisco.com, the Cisco DistributedDirector—the authoritative name-server for the www.cisco.com subdomain—responds by using DRP to query selected routers that support the Web servers in the network infrastructure for user-to-server distance information or other configurable metrics. The DistributedDirector then sorts the DRP responses it receives from the routers and,

using standard DNS, returns the IP address of the "best" server (normally defined as the closest) to the user. The client then transparently connects to this returned IP address to obtain the desired service.

HTTP Session Redirector Mode

Network administrators appreciate the ease of use offered by HTTP session redirector mode because, to operate in it, they need only add a few records to their primary domain name-servers. Applicable for HTTP services only, this mode allows network managers to configure the Cisco DistributedDirector so that it redirects HTTP sessions, such as queries sent to http://www.cisco.com, to the "best" distributed Web server. DistributedDirector accepts the initial HTTP connection, and then, using DRP, queries the routers in the network infrastructure in the same way as described for the DNS caching name-server mode. After determining the nearest distributed Web server, the system creates a new URL for that server and sends it to the user along with the HTTP status code, "302 Temporarily Moved." The user is then transparently connected to this new URL.

Using Cisco LocalDirector and Cisco DistributedDirector Together

The Cisco DistributedDirector can be used to distribute traffic among several geographically dispersed Cisco LocalDirectors. In the configuration below, the Cisco DistributedDirector directs clients to the "closest" Cisco LocalDirector. The Cisco LocalDirectors then perform local, intelligent load-balancing to ensure that connections are allocated to the servers with the highest availability at each distributed server site. This configuration, recommended by Cisco for these complementary technologies, provides both local scalability for redundant clustered servers and global scalability for geographically dispersed Internet service sites. ▲▲

Cisco LocalDirector and DistributedDirector—Complementary Solutions