HOW LOAD BALANCING STEERS WEB TRAFFIC TO THE PROPER SERVER

Information in data packets triggers Web switches to tap servers that hold the content requested

1. A user types in a URL or clicks on a link containing a Web site address.

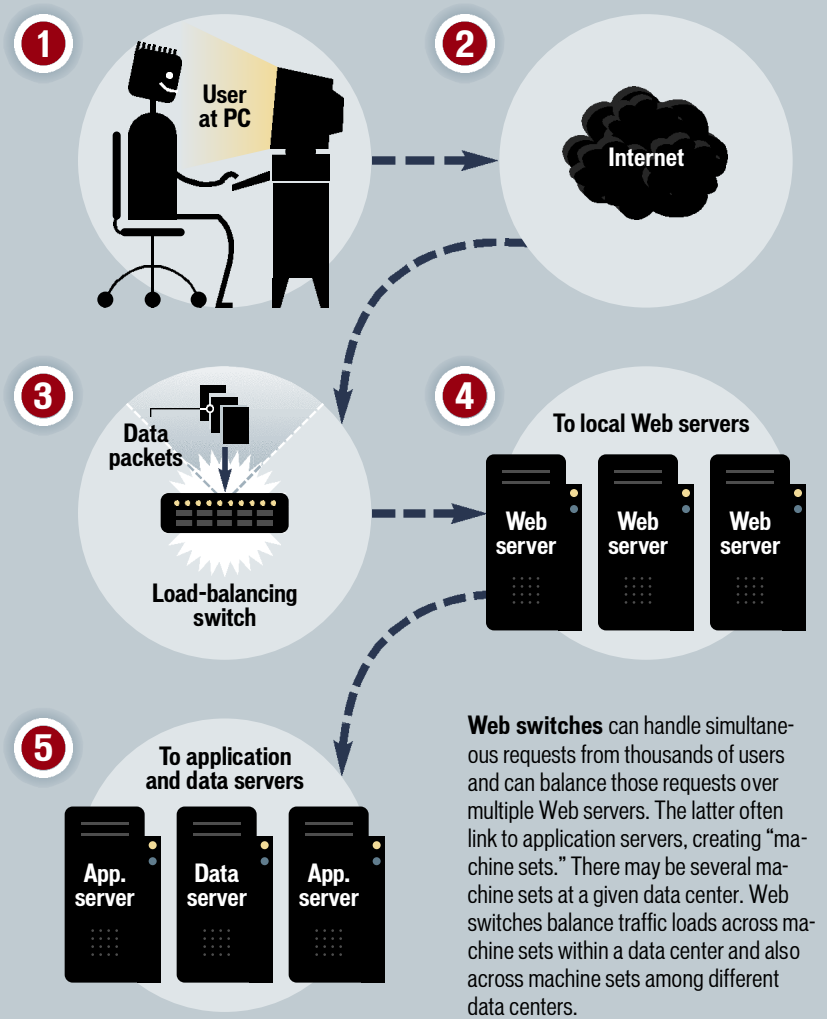
2. The browser in the user's PC (the client) issues a request through the Internet for a connection with the Web site specified by the address (the host).

3. A Web switch (load balancer) accepts the connection on behalf of the resources at the Web hosting site. The Web switch then signals the client through the Internet that a connection between client and host has been established.

4. The Web switch reads information carried by data packets about the content that the user is requesting – for example, what content is being requested, where the content is located and whether or not the individual requesting it has been approved to access the content. The Web switch also can determine the quickest path between the client and server, and route requests to another data center when appropriate – or even fulfill requests from multiple servers across several data centers.

5. Using information it has gleaned about the content requested, the Web switch routes the request to the server best able to respond quickly and fully to the user's request. It detects servers that have failed or are slow and routes traffic to servers that are more available.

Web switches examine data packets and route them according to user needs



Web switches can handle simultaneous requests from thousands of users and can balance those requests over multiple Web servers. The latter often link to application servers, creating "machine sets." There may be several machine sets at a given data center. Web switches balance traffic loads across machine sets within a data center and also across machine sets among different data centers.