Location-Aware Load Balancing

Lori MacVittie, 2010-07-07

No, it's not global server load balancing or GeoLocation. It's something more… because knowing location is only half the battle and the other half requires the ability to make on-demand decisions based on context.

In most cases today, global application delivery bases the decision on which location should service a given client based on the location of the user, availability of the application at each deployment location and, if the user is lucky, some form
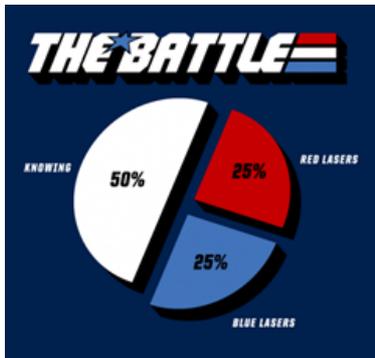
of performance-related service-level agreement. With the advent of concepts like cloud bursting and migratory applications that can be deployed at any number of locations at any given time based on demand, the ability to determine not just the user location **accurately** but the physical location of the application as well is becoming increasingly important to address concerns regarding regulatory compliance.

Making the equation more difficult is that these regulations vary from country to country and the focus of each varies greatly. In the European Union the focus is on privacy for the consumer, while in the United States the primary focus is on a combination of application location (export laws) and user location (access restrictions). These issues become problematic for not just application providers who want to tap into the global market, but for organizations whose employee and customer base span the globe.

Many of the benefits of cloud computing are based on the ability to tap into cloud providers' inexpensive resources not just at any time its needed for capacity (cloud bursting) but at any time that costs can be minimized (cloud balancing). These benefits are appealing, but can quickly run organizations afoul of regulations governing data and application location.

In order to maximize benefits and maintain compliance with regulations relating to the physical location of data and applications *and* ensure availability and performance levels are acceptable to both the organization and the end-user, some level of awareness must be present in the application delivery architecture.

Awareness of location provides a flexible application delivery infrastructure with the ability to make on-demand decisions regarding where to route any given application request based on all the variables required; based on the *context*. Because of the flexible nature of deployment (or at least the presumed flexibility of application deployment) it would be a poor choice to hard-code such decisions so that users in location X are always directed to the application at location Y. Real-time performance and availability data must also be taken into consideration, as well as capacity of each location.