Some requests are more equal than others

Lori MacVittie, 2008-19-05

This is an interesting little article on load balancing that's very close and yet very far from being completely accurate in today's world. Overall the author does a good job of hitting upon the basic concepts of load balancing, why it's important, and what some of the benefits are. But there's just one thing that I absolutely must address.

> Even distribution?
>
> Load balancing is the **even** distribution of computer processing and communication activities so that a server is not overwhelmed. Load balancing is especially important for networks where it is difficult to predict the number of requests that will be issued to a server.

Load balancing is certainly about distribution of compute resources, but today it is not necessarily about the ***even*** distribution amongst servers. That may be the definition, but it's not the way in which it is implemented in today's data-driven, object heavy web application world. In fact, I'd be willing to say that the concept of load balancing being used to evenly distribute load across servers went out in the late 1990s, when we discovered that ***even*** distribution of requests did not equate to equal performance or even equal load distribution.

That's because not all requests consume the same amount of computing resources to fulfill. A request for 10KB image does not consume the same resources as a request that requires database access to fulfill, or that requires intensive mathematical computations. Not all requests are equal, so why would you treat them equally when load balancing them?

This is why simple load balancing has not been, for many years, an optimal solution for scaling web sites and applications. Simply routing requests to servers without understanding the impact that request will have on the server in terms of resource consumption can - and often does - result in an inefficient infrastructure incapable of handling the capacity required.

Now you certainly ***can*** load balance servers evenly using round robin algorithms today if you want to, and an application delivery controller will certainly provide this functionality. But why would you ***want*** to do that when you have many more options and can ensure that not only is the load balanced, but that resources are being consumed appropriately so that your entire infrastructure is more efficient and performs much better?

This is why load balancers have evolved into ***application delivery controllers.*** It is the fundamental impetus behind introducing application fluency into load balancing and why it makes application delivery more efficient than simple load balancing.

For a better look at load balancing and application delivery, take a gander at these white papers.

Load Balancing 101: Nuts and Bolts

Imbibing: Coffee

F5 Networks, Inc. | 401 Elliot Avenue West, Seattle, WA 98119 | 888-882-4447 | f5.com

F5 Networks, Inc.
Corporate Headquarters
info@f5.com

F5 Networks
Asia-Pacific
apacinfo@f5.com

F5 Networks Ltd.
Europe/Middle-East/Africa
emeainfo@f5.com

F5 Networks
Japan K.K.
f5j-info@f5.com