

# DATABASE

## TRENDS AND APPLICATIONS

Solutions for the Information Project Team • [www.dbta.com](http://www.dbta.com)

Volume 21, Number 2 • February 2007

## Database Load Balancing Delivers Performance on Low-Cost Servers

By **Albert Lee**

The era of “super-sized systems” is winding down. Companies are starting to slim down their infrastructure, migrating towards virtualization and server farms. These lower-cost commoditized systems are more efficient, utilize resources more effectively, and squeeze all the performance and capacity possible from the infrastructure. However, until now, ensuring scalable performance, continuous availability, and data consistency for transactional databases was nearly impossible to engineer. To get around this hurdle, companies are turning toward database load balancing solutions.

### Data Tier Challenge

Throughout the 1980s, database companies pumped out big, feature-rich platforms designed for the enterprise market. End-users deployed large infrastructures - or “super-sized systems” - to host their data. This cost was a necessity organizations absorbed to guarantee performance and reliability.

Starting in the late 1990s and early 2000s, technology advances made grids, utility computing and virtualization a reality in the presentation and application layers. Companies saved millions of dollars by making these tiers more efficient, squeezing every last drop of performance out of their resources, and building an on-demand architecture that allowed them to provision resources when the need dictated.

Efficiencies in the data layer, however, remained elusive. Ensuring data consistency across multiple servers became a hurdle. The impact on performance was too much to make the architectural shift. Vendors and developers tried two-phase commit and syn-

chronous transaction replication technologies, and master-slave architectures, but all had performance or availability drawbacks.

As a result, while the application and presentation tiers grew horizontally to accommodate a more scalable on-demand architecture, the data tier remained vertical - one big, dedicated database server per application. Data was further homogenized through replication practices that required triple the infrastructure - a second server in production for high availability and an idle server in a disaster recovery site - for zero boost in performance. This led to issues with utilization, scalability, cost and true continuous availability.

### Data Load Balancing

Most businesses now have a standing order to roll out upgraded or new applications using lower-cost clustered commodity systems whenever possible. In addition, there is a push to utilize assets more effectively, doing away with idle systems on standby in case of a disaster. The challenge is to retain all the performance and reliability of super-sized systems and keep up with disaster recovery demands with more efficient commodity components. In the database world, the solution is simple - intelligent load balancing.

With many applications, load balancing is consistently used to improve the reliability and performance of server farms for Web and application servers. While each system has its own processors and memory, the cluster performs like a single system with external networking connections acting as the interconnects. The load-balancing algorithm divides requests up and assigns them to different servers based on capacity,

speed and priority. Because application workloads can often be distributed in this way without affecting consistency, the presentation and application tiers can be engineered horizontally, taking advantage of easier scalability and better resource utilization with lower cost components.

Data, on the other hand, could not be divided effectively without compromising consistency. Data is too dynamic, too fluid and too valuable to be divided onto separate servers. In addition, backup and disaster recovery would be a management nightmare, not to mention compliance efforts for government, industry and corporate regulations.

However, load-balancing solutions powered by innovative algorithms for the database level have recently been developed that intelligently distribute transactions between multiple servers while maintaining data consistency. The algorithms can be designed to identify which server has the shortest queue, the least load or lowest CPU activity - all helpful in accurately predicting which server in a cluster is most likely to execute a transaction the quickest.

Key to effective load-balancing solutions is their ability to ensure data consistency for reads and writes across all nodes in the cluster. The middleware must be able to guarantee that the data it is serving to the application is accurate and is up to date. Standard transaction replication solutions cannot guarantee this level of consistency without sacrificing performance.

By enabling database systems to be engineered horizontally - rather than the vertical architecture of “super-sized systems” environments - intelligent load balancing extends the benefits of server farms to the data level.

## Performance and Efficiency

One of the more valuable benefits in deploying commoditized servers in the data tier is full utilization of all server assets. In a vertical architecture, availability is assured by deploying an idle secondary server for high availability and another in a remote disaster recovery facility. This requires a huge capital investment and subsequent resources to deploy and manage the additional infrastructure, and, after all that, continuous availability is still not a certainty. Companies using traditional replication technology to protect large systems are forced to essentially over-provision by 100 percent or more.

Horizontal architectures enable companies to forgo deploying redundant server resources that sit idly waiting for disaster

to strike. By deploying commoditized servers in the data level, companies can put their disaster recovery resources to work full-time. Server clusters are designed to automatically and transparently withstand failures in case a server in the cluster suffers downtime. Transaction processing continues unabated, and performance and availability hardly suffers.

If one server fails, load-balancing solutions automatically service data requests from another server in the farm. The seamless failover is transparent to end-users. In addition, server maintenance, including rolling out upgrades, can be scheduled during business hours without any planned downtime. The cluster can be dispersed among remote facilities.

Finally, scalability is easier and more efficient in a horizontal architecture. IT

administrators are able to seamlessly deploy additional servers, provision resources to specific departments and boost performance for key applications or employees. The era when companies over-provisioned storage and server resources to ensure availability and performance is fading fast. New intelligent load-balancing solutions enable companies to deploy horizontal architectures in the data tier, improving asset utilization, enabling on-demand capacity, and ensuring continuous availability without sacrificing performance.

*Albert Lee is the CEO of xkoto, a provider of load-balancing solutions, and, with IBM, gives customers the means to deploy on-demand database systems. [www.xkoto.com](http://www.xkoto.com)*