Delivering High Availability for Inktomi Search Engines

Eric A. Brewer
Chief Scientist, Inktomi Corporation
Professor, UC Berkeley
brewer@full-sail.CS.Berkeley.EDU

Inktomi provides the back-end for several well-known search engines, including Wired's HotBot and Microsoft's MS Start page. The services are supported by a highly available cluster with more than 300 CPUs and several hundred disks.

After a long evolution starting with a traditional RAID-based approach to availability, we now use an extremely low-cost software-only approach, with very little replicated hardware.

The first insight is to be Machiavellian about precisely what "highly available" means; i.e., what are tolerable failure modes? Although we have in the past provided the traditional expectation, "all the data, all the time", we now provide a weaker (but far cheaper) promise:

"some of the data all of time, and (probabilistically) all of the data most of the time". The latter provides linear degradation in document availability based on the number of concurrent failures, with single failures affecting less than 1% of the database.

The second trick is to exploit the fact that search engines (and most other web-based databases) are totally dominated by reads rather than writes. We take this notion to an extreme by building a custom database with essentially no locks and one simple atomic operation, which is an atomic swap of a contiguous set of records. In a clustered environment, this one operation allows us to replace atomically all of the following with out taking down the service: the database (part or whole), the OS, disks, CPUs, power, networking cards, etc. We have even used this technique to physically move the entire cluster 60 miles without even a minute of downtime.

Finally, we will talk about many practical issues that we have had to solve to reach quality uptimes statistics in practice.

SIGMOD '98 Seattle, WA, USA
© 1998 ACM 0-89791-995-5/98/006...$5.00