CUSTOMER SOLUTIONS

Featured Product

Berkeley DB High Availability

Google is the world's most popular Web search engine. To provide fast response to user queries, the company maintains a large index of Internet data that includes billions of Web pages, images, and documents. Every day, Google's search engine responds to hundreds of millions of requests and provides results in more than 100 languages and dialects.

www.google.com

SLEEPYCAT SOFTWARE
Makers of Berkeley DB

Google Accounts

Building on the success of its search engine and advertising business model, Google has undertaken a growing number of user-oriented services that involve personalization. These include Gmail, Google Alerts, Google Groups, Google Personalized Homepage, Froogle Shopping List, as well as Google Answers and Personalized Search, all of which use a unified sign-on service called Google Accounts. Behind the login process sits a highly distributed database from Sleepycat Software, which manages user settings.

To provide these many services with a common user authentication mechanism, Google began doing what it had always done: developing critical infrastructure in-house. The single sign-on mechanism needed to be highly scalable, highly available, very fast, and tailored to Google's unique hardware configuration. The company's initial designs focused on a custom-built database supplemented with standard replication algorithms.

During early development of the sign-on mechanism, however, a Google software engineer who was familiar with Sleepycat Software suggested that Google test a just-released, high-availability version of Sleepycat's Berkeley DB. This product, called Berkeley DB High Availability (HA), is a database engine with carrier-grade scalability and reliability driven by a high-performance replication design and automatic failover.

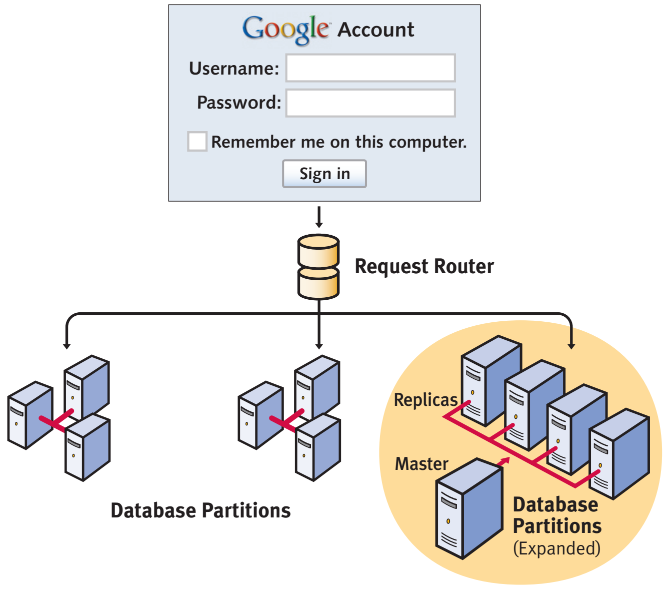
After a comprehensive evaluation period, Google added Berkeley DB HA to its infrastructure. In addition, the company deployed the replication design mechanisms provided by Berkeley DB HA. Today, these Sleepycat technologies handle user settings for all logins to Google Accounts.

Google Accounts Infrastructure

The Google Accounts software must enable user authentication with low latency, even under heavy loads. The authentication step is a transactional event that requires fast, reliable, scalable, persistence and robust high-availability. To deliver this level of service, Google Accounts uses Berkeley DB HA for the storage and retrieval of user data and for replication, thereby ensuring scalability and availability.

The Berkeley DB HA database consists of key-value pairs. When users log in, their identities serve as the key to their user settings stored in Berkeley DB HA. This database is partitioned into several large blocks, each of which is a separate Berkeley DB database in its own right. A traffic management layer implemented by Google routes incoming requests to the correct partitions. (See Figure 1)

Figure 1 **How a Google Account login request is routed.**



Read requests—such as when an existing user returns and logs in—are routed to any device in the partition, be it a master or a replica. However, updates to the database—such as the addition of a new user or modification of an existing user's data—are routed only to the partition master. That master system records the change and then propagates it to the replicas. It considers the update complete when more than half the replicas have recorded the change. This design ensures that if a system failure occurs, the nodes can replicate the change among themselves and throughout the rest of the partition.

Unlike standard server-based configurations, this design has no single point of failure. This distributed philosophy permeates the Google architecture, and it requires that the software layer be designed to handle all events that can result in service interruption.

One possible failure that must be handled correctly occurs when a partition master becomes inoperable due, for example, to a hardware or network glitch. In such a scenario, the Berkeley DB HA software holds an election among the replicas, and one of those replicas is designated the new master. All remaining replicas then automatically synchronize to the new master, and processing continues without interruption. When the original master comes back online, it becomes another replica and is brought up-to-date automatically. Through this mechanism, Berkeley DB HA ensures high availability with seamless transitions if a failure occurs.

Inside Berkeley DB High Availability

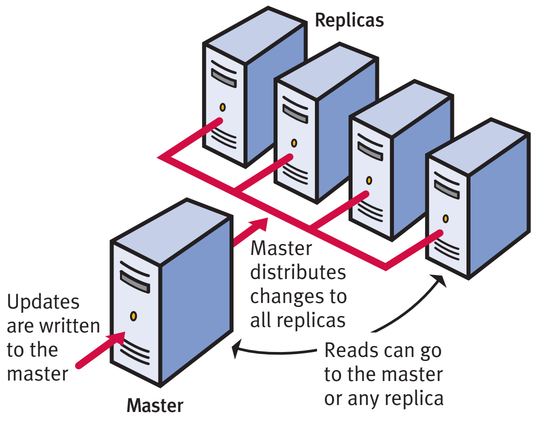
Berkeley DB is the most widely used developer database in the world, and it can be found in a wide range of open source and commercial applications. As a developer database, Berkeley DB is designed for use by programmers who need a high degree of control and flexibility. The product eschews the relational, client-server model, which imposes high processing overhead and complexity. Instead Berkeley DB uses a simple key-value pair design, on top of which programmers can implement a data representation that is most natural to the application. The key-value pair design supports true primary indices as well as secondary indices.

Packaged as a library, not a server, Berkeley DB runs in-process with the application, providing fast, local persistence. All database administration is performed programmatically, so the final application may operate without human administration or maintenance. Berkeley DB provides full atomicity, consistency, isolation, and durability (ACID) transactions and recovery, and allows the programmer to choose the level of transactional guarantees. Berkeley DB offers application programming interfaces (APIs) for C, C++, Java, Perl, Python, PHP, Ruby, and many other programming languages. Despite this, the product has a remarkably small footprint, consuming less than 500KB of memory.

Berkeley DB also provides fault resilience: If an application ends abnormally or the system is brought down for maintenance, Berkeley DB restarts immediately.

Berkeley DB HA offers fault tolerance and high availability through the single-master replication design discussed previously and illustrated in Figure 2. All updates go to a designated master node. That master distributes changes automatically to as many replicas as the application requires. Reads can go to the master or any replica. New replicas can join the group at any time, so scaling systems incrementally is accomplished easily.

Figure 2 **How Berkeley DB HA routes read and write requests.**


Replicas
Master distributes changes to all replicas
Updates are written to the master
Reads can go to the master or any replica
Master

Berkeley DB is highly configurable, and developers can select desired features to suit a specific implementation. For example, developers can use standard network protocols such as TCP/IP or platform-specific protocols to communicate among members of a group. The replication mechanism is also highly configurable, so that applications are not constrained by narrow feature sets.

Getting Your Own Copy

Berkeley DB is open source and made available under a dual license. The full software, including source code, test suite, and documentation, can be downloaded from the Sleepycat Web site.

This download provides an opportunity to see why mission-critical systems at Amazon.com, AOL, Cisco Systems, EMC, Sun Microsystems, leading stock exchanges, and other top companies use Sleepycat's Berkeley DB. To get your own copy of the same technology that these companies and Google rely on for high-performance, highly reliable data access, go to www.sleepycat.com and download Berkeley DB today.

Sleepycat Software makes Berkeley DB, the most widely used open source developer database in the world with over 200 million deployments. Customers such as Amazon.com, AOL, Cisco Systems, EMC, Google, Hitachi, HP, Motorola, RSA Security, Sun Microsystems, TIBCO and VERITAS also rely on Berkeley DB for fast, scalable, reliable and cost-effective data management for their mission-critical applications. Profitable since it was founded in 1996, Sleepycat is a privately held company with offices in California, Massachusetts and the United Kingdom.

For further information, please contact Sleepycat by sending email to info@sleepycat.com or visiting Sleepycat's website at www.sleepycat.com

SLEEPYCAT SOFTWARE
Makers of Berkeley DB

Corporate Headquarters
Sleepycat Software Inc.
118 Tower Road
Lincoln, MA 01773
USA

West Coast Office
Sleepycat Software Inc.
5858 Horton St. Suite 265
Emeryville, CA 94608
USA

European Office
Sleepycat Europe Ltd.
Coronation House,
Guildford Rd.
Woking, GU22 7QD
United Kingdom

Telephone
+1-978-897-6487
+1-877-SLEEPYCAT
(Toll-free, USA only)