Sprint: High-Performance Dependable Computing

Demand for high performance and high availability combined with plummeting hardware prices has led to the widespread emergence of large computing clusters. Such environments offer great potential for highly efficient and reliable data management systems. But realizing this vision requires revisiting several fundamental assumptions about current data management systems. This is the goal of Sprint, a project conducted at the Faculty of Informatics, University of Lugano, and funded by the Hasler Foundation and the Swiss National Science Foundation.

In recent years, web applications have become commonplace in most online businesses. These systems require fast interactive response times and the ability to serve a large number of clients uninterruptedly. At the core of a web application lies a database system responsible for storing the application's state. As systems grow in number of clients, the database invariably becomes a performance bottleneck. This happens because it is usually much simpler to replicate the other components of a web-based application (e.g., web server) than the database. Moreover, as other components are replicated, the database becomes a single point of failure, that is, the availability of the compound depends on the database ability to tolerate failures. Current solutions to increase the performance and the availability of database systems usually rely on specialized hardware and are expensive.

The Sprint project intends to design and implement a data management platform for web applications. The platform will benefit from current trends on off-the-shelf hardware components. Many commodity systems now boast main memory capacities that rival that of hard disks from a few years ago. In addition, commodity hardware can be arranged in very powerful clusters. These clusters are assembled from readily available parts and powered by free software. However, architectures of this sort demand novel approaches to the way software systems are designed. Consider for example the discrepancy in disk capacities and disk transfer rates. While disk capacities continue to double each year, access times are improving at the rate of 10 percent a year. This creates a bottleneck limiting the system's ability to access and process data, even if processors get faster. One solution to this dilemma is to focus on data management mechanisms that rely on main memory only (i.e., main memory databases), avoiding the disk overload.

A typical Sprint environment (copyright Virginia Tech).

Besides performance and availability issues, data management systems are also required to cope with the complexity of large clustered architectures. Ideally, the system should be able to adapt automatically to environmental changes. Adaptation does not only simplify cluster management but can also contribute to its availability and performance. If a critical server fails, another one takes over, increasing the availability of the ensemble. Adapting to workload variations has a more subtle advantage. If most applications only query the database, performance can be improved by increasing the number of database replicas and running transactions in parallel in different servers. But if update operations prevail, more replicas means greater overheads to synchronize and keep them up to date. The optimal degree of replication can only be achieved by systems able to adapt to the workload on the fly. Solving such problems is the prime objective of the Sprint project.

The Sprint approach

Sprint aggregates servers in a cluster and makes use of high performance techniques to overcome the latency limitations of traditional disk-based databases. Data partitioning and replication are combined for performance and high availability. Each server runs a local *main-memory database system* (MMDB). This means that the whole database state resides in the main memory of servers, although no single server is expected to contain the whole database image. As a consequence, transactions that only read data from the database depend solely on main memory and network latencies, which are orders of magnitude inferior to disk latencies. For recovery reasons, transactions that modify the database state have to write to disk. Sprint handles such cases by completely serializing disk access, which is much faster than random disk access. By removing long disk delays, both throughput (i.e., the number of transactions executed per time unit), and response time are improved.

Replication is used to increase both availability and performance. Without replication, data stored in a server that fails would become unavailable. Since every data item is replicated, clients do not notice interruptions in the service when a server fails, since data can be fetched from another server. Replication also helps improve performance: if some data items are mostly read and rarely modified, then replicating them in several servers allows simultaneous execution of transactions. Data partitioning is used to increase performance. The idea is that if portions of the database are accessed frequently, then they can be placed in different servers, so that transactions can be executed in parallel.

The research team at USI. From left to right: Lasaro Camargos, Fernando Pedone, Vaide Zuikeviciute, Rim Moussa, Marcin Wieloch and Rodrigo Schmidt.

Sprint will contribute to research on the design and implementation of future highly efficient and available adaptive data-management systems. It will revisit traditional data management algorithms from the perspective of emerging cluster technologies and seek to understand how their performance is affected by these technologies. Demand for powerful and robust data management systems will continue to increase in future years. Achieving the level of scalability and availability necessary to cope with this demand requires techniques that fully exploit modern cluster technologies and automatically adapt to environmental changes. Sprint is an open-source project. Users will be able to download it freely and use it according to their needs.

Wide-area data management

Sprint targets large computer clusters, but there is another emerging trend on future data-management systems: collections of autonomous clusters spread over large geographical areas and interconnected through wide-area links. Such architectures are designed to tolerate disaster failures (e.g., an earthquake, which could bring down a whole cluster). One of the main challenges is to reduce the large communication latencies caused by the distance between clusters. The research team at USI is investigating such aspects in the context of GORDA, a project funded by the European Union. GORDA will connect several database clusters spread across Europe. The consortium groups together academic and industrial partners, such as MySQL.

For further details:

Prof. Fernando Pedone
Faculty of Informatics
Università della Svizzera Italiana
Phone: +41 58 666 46 95
E-mail: fernando.pedone@unisi.ch
http://www.inf.unisi.ch/pedone

Web addresses:

http://www.inf.unisi.ch
http://www.inf.unisi.ch/sprint

I M P R E S S U M

Nr. 16 - March 2005
www.ticinoricerca.ch/progettomese.htm

Servizio ricerca USI - SUPSI
via Lambertenghi 10a, CH-6904 Lugano
Tel. +41 58 666 48 18; e-mail: sr@unisi.ch