Report on the 7th Workshop on Distributed Data and Structures (WDAS 2006)

Thomas Schwarz, S.J.
Department of Computer Engineering
Santa Clara University
Santa Clara, CA 95053, USA
tjschwarz@scu.edu

Mark Manasse
Microsoft Research, Silicon Valley
1065 La Avenida
Mountain View, CA, 94043, USA
manasse@microsoft.com

The seventh Workshop on Distributed Data and Structures (WDAS) took place on the campus of Santa Clara University on January 4 and 5, 2006 and drew participants actively working in research on distributed data, structures, and their applications. WDAS aims to stimulate the exchange of ideas and to be a forum for work in progress. The electronic version of the Proceedings contains all papers accepted at the conference [WDAS]. In addition, the workshop presentations helped to select a subset that will appear revised in the Proceedings in Informatics Series at Carleton Scientific. Two papers could not be presented by authors because of the difficulties of obtaining an entry visa to the United States, unfortunately a more and more common phenomenon. We now give an overview of the topics discussed at the workshop and the related presentations.

1 Distributed Storage Systems

The workshop had two keynote addresses. Darrell Long (UC Santa Cruz) laid out the reliability challenges facing large-scale storage systems. Institutions like those at the national laboratories now face demands for petabytes of highly available data using commodity disk drives. Once the installation reaches a certain scale, device loss and malfunction become normal occurrences and simple strategies such as RAID Level 1 or 5 are no longer sufficient to make data loss a rare occurrence. The challenge of better manageability needs to extend to failure detection and failure recovery. The resulting research problems need to be addressed by both academic research with its quest for innovation and by commercial research with its financial resources and product orientation.

Bisson, Wu, and Brandt deal with another problem of very large storage sites, namely energy consumption. The disk based storage system of the future might try to turn disks off as much as possible. As a trend, disks become smarter and the object storage devices of the near future will combine a small CPU with a network card and a disk drive and possibly a small amount of non-volatile memory such as flash or MRAM. Such a disk can react to its environment while the disk itself is powered off. Ceph, the object storage device file system project at UC Santa Cruz exploits these features. In the paper presented at the workshop, the authors improve of the state of the art by including knowledge of the power status of other disks into the decision process and by redirecting write requests from devices in low power mode.

Currently, metadata is not used extensively across applications, though individual applications that manage large music or photo collections maintain a rich set of metadata specific to the type of data and the application itself. Another project at the Storage Systems Research Center at UC Santa Cruz called *Graffiti* (Bobb, Eads, Storer, Brandt, Miller, Maltzahn) tries to investigate ways to make metadata more portable and more exploitable by adding a new distributed metadata layer above the operating system. In essence, a new metadata layer will make the task of application programmers simpler and allow sharing of metadata across three dimensions: applications, people, and computers.

2 Video-On-Demand

Video on demand system have high bandwidth requirements. Proactive or broadcasting protocols schedule streams at certain times and reduce bandwidth by sending the same video stream to more than one user. Reactive protocols use techniques such as batching and piggybacking; they can also force users to participate by sending a just received stream to other users (tapping or chaining). Pâris presented such a protocol that allows for different capacities between upload and download and maximizes the contribution of near-simultaneous consumers of the same video.

3 Traditional Data Structures

Otoo presented work on an extendible array data structure that was motivated by needs in scientific computing to change the range of a dimension of a multi-dimensional array while maintaining good performance.

Erlingson, Manasse, and McSherry use multiple hash functions to improve traditional, static hashing (an improvement on "Cuckoo Hashing" by Pagh and Rodler, Fotakis *et al.* and Panigrahy). Their technique can achieve load factors of 99.9% in 99% of all instances. This paper generated the longest discussions on how best to apply its techniques to distributed systems. One of the exciting implications of this work is the possibility of types of distributed hash tables using overlay networks. One should be able also to optimize the mappings of the bucket structures of distributed hash tables.

4 P2P

P2P databases or P2P information retrieval need to partition a large set of nodes into k partitions with additional requirements such as load-balancing without global knowledge. Bickson, Dolev, Weiss, Aberer and Hauswirth apply probabilistic graph models in the context of P2P systems for two problems. First, they count the number of nodes of a specific type without requiring a specific communication topology. Second, they perform a distributed graph coloring. Both algorithms are implemented using a belief propagation inference algorithm and exhibit astonishingly good accuracy and low overhead.

Martins, Pacitti, and Valduriez address the problem of replica consistency under more dynamic conditions on semantic reconciliation than the classical ones. These conditions prevail in P2P, grid and mobile computing systems. Users might perform updates, join, and leave the network whenever they wish. Existing semantic reconciliation solutions are typically performed at a single node and are inappropriate for such a dynamic environment. Starting from IceCube, the authors propose DSR, which enables optimistic multi-master replication and assures eventual consistency among replicas.

5 SDDS

Scalable Distributed Data Structures allow access to distributed data in times that are largely or even completely (LH*) independent of the number of storage sites and that do not require a central coordinator for data addressing. Introduction of dynamic data structures such as linear hashing and B-trees did away with static techniques such as static hashing and ISAM. As the industry starts dealing with large database tables that need to be distributed over several sites, these distributed tables are still administered in a static fashion. The Ph.D. work of Soror Sahri under Litwin presented a prototype called SD-SQL server that is directed towards remedying this situation. It uses SDDS principles to distribute a growing table without administrator intervention over more and more sites (currently up to 250). The user manipulates the tables through a dynamically adjusted and updatable distributed partitioned view. Queries are processed in parallel at all storage sites and response time benefits from the distribution.

The Ph.D. work of Damian Cieslicki under Schwarz is devoted to providing growing SDDS with a sub-layer providing scalable high availability in a generic manner. Availability is achieved by using erasure correcting codes (m/n codes) to create parity data with which data on lost nodes can be reconstructed. His work should improve considerably the update performance of LH*$_{RS}$ (Litwin *et al.,* TODS, Sept. 2005).

6 Service Oriented Computing

Gavran, Milanović, and Srblić deal with the interesting problem of developing large-scale distributed applications based on Service-Oriented Computing. They developed a "simple service composition" programming language. It allows for fast and simple design and implementation of applications in the service-oriented programming model.

7 Web Spam

Marc Najork from Microsoft Research – Silicon Valley gave a second keynote address in which he discussed heuristics for detecting spam web pages, a burden imposed on any search engine by manipulating content providers. His interesting presentation was based on his joint work with Fetterly and Manasse [FMN04].

Acknowledgments

The conference was sponsored by the Microsoft Research Silicon Valley (Mountain View, CA) and by the host, the School of Engineering at Santa Clara University. C. Maltzahn, W. Litwin, J. F. Paris, and P. Valduriez actively helped with this write-up.

Papers at WDAS 2006

J.F. Pâris: Using available client bandwidth to reduce the distribution costs of video-on-demand services.

Danny Bickson, Danny Dolev, Yair Weiss, Karl Aberer, Manfred Hauswirth: Indexing data-oriented overlay networks under belief propagation.

Ivan Gavran, Andro Milanović, Siniša Srblić: End-user programming languages for semantic reconciliation.

Vidal Martins, Esther Pacitti, Patrick Valduriez: A dynamic distributed algorithm for semantic reconciliation.

Ulfar Erlingson, Mark Manasse, Frank McSherry: A cool and practical alternative to traditional hash tables.

Timothy Bisson, Joel Wu, Scott Brandt: A distributed spin down algorithm for an object-based storage device with write redirection.

Witold Litwin, Soror Sahri, Thomas Schwarz: Prototyping SD-SQL server: a scalable distributed database system.

Damian Cieslicki, Stefan Schäckeler, Thomas Schwarz: Highly Available Distributed RAM (HADRAM): Scalable availability for scalable distributed data structures.

Nikhil Bobb, Damian Eads, Mark Storer, Scott Brandt, Carlos Maltzahn, Ethan Miller: Graffitti: A framework for testing collaborative distributed metadata.

Syed Ahsan and Abad Shah: Biological databanks: distribution, heterogeneity, diversity, and provenance.

Ekow Otoo: Parallel and distributed access of dense multi-dimensional extendible array files.

References

[FMN04] Dennis Fetterly, Mark Manasse, and Marc Najork. Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages. 7th International Workshop on the Web and Databases (June 2004)

[WDAS] The WDAS proceedings are available in electronic form at www/soe.ucsc.edu/wdas06.