Proceedings of the Workshop on Fault-Tolerance in Parallel and Distributed Computing, San Diego, CA, pages 58–59, 1987

The Management of Consistency in Fault-Tolerant File Systems

Jehan-François Pâris
Darrell D. E. Long
Walter A. Burkhard

Computer Systems Research Group
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, California 92093

Critical data are often replicated in distributed systems to protect them against site failures and network partitions or to allow higher access rates. This trend benefits from recent advances in computer network technology and reductions in the cost of storage media, which have made the replication of important files on several sites a cost-effective proposition.

The existence of several copies of the same file raises the issue of *file consistency*. It would be quite unreasonable to require the system users to be responsible for the consistency of data stored at multiple sites of a network. Distributed file systems implementing file replication need to enforce an access policy isolating their users from that issue and maintaining a consistent view of all replicated data under all possible circumstances. Several consistency protocols have been discussed in the literature, including token based schemes, active copy schemes and schemes based on quorum consensus. We still lack however a comprehensive study of the costs and performance of these consistency protocols in terms of their traffic overhead and on the restrictions they impose on file availability.

We have developed in recent years a replicated file system prototype to study these issues and to investigate the feasibility of more efficient consistency protocols. Our Gemini prototype [BMP87] runs on a network of **UNIX** machines linked by a subnet using the DARPA TCP/IP transport layer protocol. Because of its extremely modular conception, it provides an ideal test-bed for the study of consistency protocols. This project has already given birth to several novel consistency protocols including Dynamic Voting [DaBu85], Voting with Witnesses [Pari86a, Pari86b], Naive Available Copy [LoPa87, CLP87], Optimistic Dynamic Voting and Topological Dynamic Voting [PaLo88].

In our talk, we will briefly review these protocols and present the tools used to evaluate their performance [CLP87, PLG88] with respect to other consistency schemes such as Majority Consensus Voting [Elli77, Giff79] Available Copy [Good83], Lexicographic Dynamic Voting [Jajo87] and Regeneration [PNP86]. Our presentation will focus on two key issues. First, copy locations and network topology are to play a primary role in the selection of a consistency protocol for replicated data. Second, there is no need to compromise between file availability and network traffic overhead as all the protocols can be efficiently implemented by allowing them to use possibly out-of-date information.

References

[BeGo84] P. A. Bernstein and N. Goodman, “An Algorithm for Concurrency Control and Recovery in Replicated Distributed Databases.” *ACM Trans. on Database Systems*, Vol. 9, No. 4 (Dec. 1984), 596-615.

[BMP87] W. A. Burkhard, B. E. Martin and J.-F. Paris, “The *Gemin*i Fault-Tolerant File System: the Management of Replicated Files.” *Proc. Third International Conference on Data Engineering*, Los Angeles, Calif. (February 1987), pp. 441-448.

[CLP87] J. L. Carroll, D. Long and J.-F. Paris, “Block-Level Consistency of Replicated Files.” *Proc. Seventh International Conference on Distributed Computing Systems*, Berlin, Germany (Sept. 1987), pp. 146-153.

[DaBu85] D. Davcev and W.A. Burkhard, “Consistency and Recovery Control for Replicated Files.” *Proc. 10th ACM Symposium on Operating System Principles*, (1985) pp. 87-96.

[Elli77] C. A. Ellis, “Consistency and Correctness of Duplicate Database Systems.” *Operating Systems Review*, 1, 1977.

[Giff79] D. K. Gifford, “Weighted Voting for Replicated Data.” *Proc. Seventh ACM Symposium on Operating System Principles*, 1979, pp. 150-161.

[Good83] N. Goodman, D. Skeen, A. Chan, U. Dayal, R. Fox and D. Ries, “A Recovery Algorithm for a Distributed Database System.” *Proc. Second ACM Symposium on Principles of Database Systems*, Atlanta, Georgia (March 1983), pp. 8-15.

[Jajo87] S. Jajodia, “Managing Replicated Files in Partitioned Distributed Database Systems.” *Proc. Third International Conference on Data Eng*ineering, Los Angeles, Calif. (February 1987), pp. 412-418.

[LoPa87] D. Long and J.-F. Paris, “On Improving the Availability of Replicated Files.” *Proc. Sixth Symposium on Reliability in Distributed Software and Database Systems*, Williamsburg, Virg. (March 1987), pp. 77-83.

[Pari86a] J.-F. Pâris, “Voting with Witnesses: A Consistency Scheme for Replicated Files.” *Proc. Sixth International Conference on Distributed Computing Systems*, Cambridge, Mass. (May 1986), pp. 606-612.

[Pari86b] J.-F. Pâris, “Voting with a Variable Number of Copies.” *Proc. Sixteenth Fault-Tolerant Computing Symposium*, Vienna, Austria, (July 1986), pp. 50-55.

[PaLo88] J.-F. Pâris and D.E. Long, “Efficient Dynamic Voting Algorithms.” *Proc. Fourth International Conference on Data Engineering*, Los Angeles, Calif. (February 1988), to appear.

[PLG88] J.-F. Pâris, D.E. Long and A. Glockner, “A Realistic Evaluation of Consistency Algorithms for Replicated Files.” *Proc. Twenty-first Annual Simulation Symposium*, Tampa, Fla. (March 1988), to appear.

[PNP86] C. Pu, J. D. Noe and A. Proudfoot, “Regeneration of Replicated Objects: A Technique and its Eden Implementation.” *Proc. Second International Conference on Data Engineering, Los Angeles*, Calif. (February 1986), pp. 175-187.