



## Distributed Database/File Systems

### Introduction

John C. Peck

Computer Science Department

Clemson University

Clemson, SC

During the last fifteen years database management has grown so quickly that today it is found on machines of every size from the largest mainframes to the smallest micro computers. Almost without exception, the production information systems which use database technology store the data at single site, usually local to the main processor. The reasons for this fact are:

1. The most expensive component of a computing system has traditionally been the central processing unit. Duplicating central processors at multiple sites, until recently, was cost prohibitive.
2. Relatively slow transmission speed and resulting poor performance made distribution infeasible.
3. Algorithms and other techniques for solving the problems in 2 above have been slow in coming.
4. Commercially available software has been slow in reaching the marketplace because of 1, 2, and 3 above.

With the relatively recent advancements in price performance of new machines offered by every vendor, distribution of data to multiple sites is rapidly becoming commonplace. No longer is the cost of the central processing unit an obstacle to distributing the processing associated with distributed databases. The processing power needed for distributed database systems is already in place in most big corporations. Micro computers are being used in almost every department for word processing, spreadsheets and other local computational needs. Local area networks connecting these micro computers allow rapid access to data stored at different sites.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

Many of these machines have considerable disk storage and, unfortunately, are used to store redundant copies of corporate data to meet end user needs. File transfer utilities have provided a mechanism to create private copies of corporate data on micro computers remote from the centralized corporate database.

Although many end user problems have been solved with private copies of data, many of the problems which originally motivated the development of centralized database management technology have now reappeared. Remote users now maintain and update their local files so that inconsistency of data found in multiple files is again a problem. Standards which were enforced through the use of tools like data dictionaries are now easily violated as each user can now define new data elements and new codes at will. The ability to relate information across organizational lines is sometimes impossible. Audit trails of access and update activities are often lost. Security is now in the hands of the end user and is many times compromised.

The solution to these problems is to move from redundant files to true distributed database management systems. If we are to hasten this movement then research and development efforts associated with distributed database management systems must be emphasized. Issues number 2 and 3 listed above are now the targets for much of this research. The papers in this session address three problems and associated solutions which are part of this research.

The first paper addresses a problem associated with access to a distributed database through a local area network of workstations. The characteristics of systems which automatically control the integrity of private and shared data, both original and redundant, is described.

The second paper addresses the problem of synchronization required when multiple users access common data contained within a distributed database. A model is described which assists in approximating delays due to serial access to shared distributed data resources which contain read and write locks.

The final paper in this session proposes ways in which parallel processing techniques can be applied to three file organizations. Since distributed database systems will necessarily perform parallel processing across the nodes containing fragments of the global database, the concepts described in this paper suggest ways to improve access times.