Connection of a Climate Model Database and Mass Storage Archive(s)

Michael Lautenschlager, Hannes Thiemann
Deutsches Klimarechenzentrum GmbH
Bundesstrasse 55
D−20146 Hamburg, Germany
data@dkrz.de
tel +49−40−41173−334
fax +49−40−41173−400

Abstract
An overview of an existing climate database which allows for storage of terabyte data volume is presented. Some features like the general architecture and the integration with an HSM are highlighted in more detail.

1 Introduction

The DKRZ (Deutsches KlimaRechenZentrum) is the central climate computing center for Germany. Numerical models for the coupled climate system were developed and integrated on the computer environment at DKRZ. The results are archived and disseminated for the climate research community in Germany and Europe as well as world wide. The mass storage archive contains currently 80 TByte (Aug. 99) of climate model data (90%) and of observational data (10%).

The TByte archive size is correlated with user access problems to climate model data. Data are basically accessible via a UNIX file system on the file server. No related catalogue information is available. Data are archived as time series of 4 dimensional data blocks (model raw data), whereas users access data as time series of 2 dimensional records (processed data, e.g. 2m temperature). Archived model data sets are stored on sequential storage devices, requested data are reloaded as complete files into the file server's disk cache. Then users have to transfer the files to their local client machines by FTP.

A database system, CERA (Climate and Environmental data Retrieval and Archiving system), has been developed in order to organize the data archive and to improve the users access to climate model data. Processed data and raw data are presently stored together with their description (meta data) in the CERA database [1]. Although not the entire archive is part of the database system, the currently existing CERA database with a size of 2 TByte contains more data than magnetic disks are available. Consequently parts of the database have to be stored on tapes yielding interaction problems between the database management system and the mass storage system. A hierarchical storage management including disks and tapes is presently not supported by commercial database management systems.

2 General architecture

The CERA data system itself is separated into three parts: data model [2], data hierarchy, and data processing. The CERA data hierarchy reflects three different levels of storage and access performance: metadata, processed climate data and climate model raw data. All parts of the data hierarchy are stored stored in tables of the CERA database.

The metadata contain the description of the climate data archive (data catalogue). The access should be as quick as possible.

The processed climate data contain frequently requested data structures. They are extracted and processed from the climate model raw data and are stored directly in the CERA database according to user requirements. The processed climate data can be accessed as BLOB entries (Binary Large Objects) in database tables. The processed climate data should be preferably available on hard disks near the database system in order to realize a performant access. The CERA data hierachy reflects the granularity in data warehouse architectures [3].

The third level in the data hierarchy contains the climate model raw data. Monthly accumulated model results are directly written into database tables as BLOBs. The data access is less performant than for processed climate data, because these data are stored on magnetic tapes under robot access. Only for a specific user request the raw data will be transferred from the file server into the database cache.

The basic problem with respect to storage of database files is to establish a flexible connection between database and mass storage system which allows for data migration and de−migration in dependence of user requests to the CERA database.

3 Database and Mass Storage Archive

Even the amount of data in the CERA database is too large to store the data exclusively on magnetic disks. It is only possible to store the actually used data sets on disk, the others have to be migrated to tapes of the mass storage system (MSS). The climate data in CERA are stored as BLOB‘s in database tables. Therefore the database management system (DBMS) has to interact with the mass storage system and the related archive system. At the DKRZ ORACLE is used for database management and the mass storage archive is administered by UniTree. The basic design assumption for DBMS is that all database files are randomly available on magnetic disks. A direct integration of a hierarchical storage management (HSM) is not available. The integration has to be developed individually in dependence of the used DBMS and the used archiving software.

Within ORACLE data are stored in tables and tables are summarized in tablespaces. The physical storage level is connected to the tablespaces (TS). When ORACLE is implemented on a (Unix) file system tablespaces are stored in one or more files, the database files (DBF). Tablespaces can be in different states; online, offline and read only are of interest with climate model data archiving [4].

- Online is the normal status. Data within tablespaces of this status are immediately accessible.

- Temporarily deactivated tablespaces are offline. DBFs belonging to such tablespaces are offline too. Objects residing in these tablespaces are not accessible. These tablespaces will be not opened at startup of the database.
- Read Only TS are tablespaces in which all objects can not be changed any more. The database system do not accesses these database files in write modus. This status is especially important in connection with backup and recovery. Read only tablespaces can be either online or offline.

When data are currently not needed it is possible to migrate DBFs if their affiliated tablespace is set offline. If data are requested by an application (e.g. user request) the DBF‘s have to be demigrated and the tablespace has to be set online again.

This mechanism has been automated at DKRZ by developing and installing a storage broker which acts as an interconnect between (database)−applications, the database system and the mass storage system. The storage broker is divided into interacting processes with some of them running inside the database, some of them outside using ORACLE‘s External Procedure Calls [4]. The main processes are:

- The main−storage broker accepts requests, checks space within database disk cache, allocates space and sends requests to other brokers.
- The make−space broker clears disk space based on dataset priorities in order to allow for de−migration.
- The de−migration broker de−migrates database files from mass storage system back to database disk cache.

The storage broker controls the migration and de−migration according to database requests and to that disk space which is available to the CERA database. Database request query optimization is realized in a disk cache area which is controlled by the broker and which is strictly separated from the archiving system. The migration/de−migration strategy is highly flexible. Priorities of requests are calculated online based on dataset, user and system load characteristics as well as on recorded database access statistics. These statistics may allow also for a pre−caching algorithm. A persistent interconnection between these processes allows for re−launch of requests even after database crashes.

Data extraction from and the data delivery into the mass storage system is realized by standard FTP in order to maintain independence from the archiving system.

4 Conclusions

The separation of the database migration disk cache from the standard HSM disk cache allows for an implementation of a database driven migration strategy and for ndependence from the connected HSM system. This two level cache approach is robust and provide a large degree of independence between RDBMS and HSM.

As the complete system is implemented on a Unix file system and not on raw devices standard file transfer mechanisms like ’ftp’ and ’copy’ can be used to connect the two disk cache areas. Therefore practically all HSM systems can be used.

References

[1] M. Lautenschlager and M. Reinke (Ed.), "Climate and Environmental Database Systems", pp. 197, Kluwer Academic Publishers, Boston 1997

[2] M. Lautenschlager, F. Toussaint, H. Thiemann and M. Reinke, "The CERA−2 Data Model", pp. 53, Technical Report No. 15, DKRZ, Hamburg 1998

[3] W. H. Inmon, "Building the Data Warehouse, Second Edition", pp. 401, John Wiley & Sons, Inc. 1996

[4] George Koch and Kevin Loney, "ORACLE8 − The Complete Reference", Orcale Press, pp. 1300, Osborne McGraw−Hill 1997