

Case Study

Scientific Data Management at DKRZ

Summary

Climate research helps generate scientific evidence of human induced climate change. The German Climate Research Centre (DKRZ) in Hamburg has acquired a NEC vector supercomputer for this purpose and significantly further expanded its databases using Oracle 9i on Linux. Researchers can simulate severe environmental consequences such as the thawing of glaciers, the rising of sea levels and even the collapse of the Gulf Stream with the new system, which was designed and installed by NEC High Performance Computing Europe. In addition, the centre is now optimally equipped to archive and analyze its research results with a petabyte Oracle database.

Managing vast amounts of data

In search of proof of climatic change, climate research relies on both empirical data encompassing such information as prevalent temperatures in the last centuries as well as working with models which simulate climatic changes on the basis of physical laws with the aid of mathematical equations. These simulations produce huge amounts of data and require high-performance computers such as those managed by the group of scientists at the German Climate Research Centre.

The amount of results produced and archived doubled nearly annually in the years 1999 to 2003, growing from 2 terabytes to 25 terabytes. And this growth has now been boosted even further with the new high-performance computer from NEC. The database with the simulation results already encompassed approximately 125 terabytes of data in December 2004, and is continuing to grow at leaps and bounds.

Year	Database size
1999	2 Tbytes
2000	4 Tbytes
2001	7 Tbytes
2002	11 Tbytes
2003	25 Tbytes
2004	125 Tbytes

This gigantic volume of data must be managed in such a manner that the scientists who later analyse it can access it easily despite the vast amount and complexity of information. Recently, a sophisticated hierarchical storage solution has been used for this purpose. This solution was put into operation by NEC as the main contractor in cooperation with partners and the Model and Data Group at the Max Planck Institute for Meteorology. An Oracle-9i database on Linux stores the simulation results at the German Climate Research Centre and regularly swaps out the data in a huge tape library in order to limit the required disc space.

Climate models calculate on a three-dimensional grid spanning the entire globe. The overall state of the system is usually saved for each of the grid points in time intervals of six hours of model time. The state contains a number of parameters: temperature, precipitation, wind, etc. Especially accurate climate forecasts require detailed horizontal resolution of the models. If scientists increase the resolution of the grid, enormous quantities of data are generated for each individual time interval.

Resolution	Per time interval	Per model month	Per 500-year run
110 km	100 Kbytes	5.2 Gbytes	30 Tbytes
300 km	16 Mbytes	650 Mbytes	3.7 Tbytes

Higher resolution enables the scientists to simulate both regional forecasts as well as processes changing rapidly in space and time, such as cloud formation. The adequate simulation of many feedback loops in the climate system can be better examined and incorporated into models with an increased resolution.

From 25 to 125 terabytes

A relatively traditional architecture was used at the DKRZ until early 2002. A Cray computer was linked to a Hierarchical Storage Management System (HSM) over a network. Post-processing computers were also linked over the network, as was a large Sun server used to run a relatively compact monolithic Oracle database of 25 terabytes in size.

NEC High Performance Computing Europe (HPCE) then equipped the DKRZ with the new vector computer SX-6. At this time, the latest generation of the SX series dominated the Top 500 benchmark list of the fastest supercomputers in the world. The forerunner was the Earth Simulator installed in Yokohama. With the 100-fold faster supercomputer at DKRZ the production of result data increased many times over. A new concept to manage the simulation results for this foreseeable situation likewise had to be developed.

Rapid Storage

A system was sought which could both store data on discs as fast as it is produced by the supercomputer as well as offering scientists rapid access to the archived data. It was to implement the fastest network technology and run on Linux.

At the start of the project, it soon became clear that the new database system needed to consist of several servers working together. An approach using a single symmetrical multi-processing system was thus rejected and the structure of the Oracle Federate Database chosen instead: five computers with several database instances and joint external representation of the available data. The commercially available support, good experiences with scalability, existing expertise and good data handling were all factors in favour of maintaining the tried-and-proven Oracle platform.

The NEC-TX7 systems used for the database cluster run on Linux are Itanium-2 computers. The system bus can transport 6.4 GBytes/s, thus achieving a stand-alone performance of 4.0 GFlops per CPU. Memory access takes place via Cache-Coherent Non-Uniform Memory Access (CCNUMA). The system is organised in cells with four CPUs each.

The Linux used at DKRZ is a Red Hat Advanced Server with kernel 2.4.18. The kernel was adapted for operation in the data centre with several necessary features. For example, it provides support for the Global File System (GFS). In GFS, a more advanced development of the NFS (Network File System), clients retrieve data directly from the SAN (Storage Area Network) via Fibre Channel. The GFS server ensures consistency in the file system by managing meta data such as the inode operations on the mass storage device and informing clients of the block addresses.

Hierarchies for data

The Oracle database is embedded in a Hierarchical Storage Management System (HSM) in order to store the vast amounts of data. Files are automatically copied to a back-end system based on certain criteria according to which they can also be removed from the disc. The HSM system is a Unitree/DXUL on a Linux basis. The additional product Disk Xtender handles the swapping-out. Disk Xtender uses the filesize as well as the time of the last access as criteria.

If a user accesses the swapped-out files via the Oracle database, the Disk Xtender starts a process that reloads the specific item needed in the background. Tape access usually takes approximately five minutes. During this time, the tape is selected, automatically loaded in the drive, forwarded to the desired record and read.

As most scientists are especially interested in individual variables (e.g. surface temperature), the data is also saved in the database in keeping with this structure, as made possible through the separation of result data in the form of BLOBs (Binary Large Objects) and metadata. Each time series is stored in chronological order in a single table. The data for a certain time point forms one BLOB line. Some tables very soon reach a size of 10 Gbytes or multiples thereof. Calling up time series for individual variables enables users to purposefully limit their data requests, thereby reducing the amount of data transfer. One significant advantage is the drastically reduced need for database disc cache space.

Petabyte database in sight

The new database is a great step forward into the future for the Model and Data Group at the Max Planck Institute for Meteorology. The reliability and availability of the architecture used for a distributed Oracle database have proven good in practice and offer space for at least one petabyte of data. Scientists have only to generate this data through more detailed regional resolution and researching new physical processes in the climate system.