The 3rd International Workshop on Next Generation Climate Models
for Advanced High Performance Computing Facilities, March 28-30, 2001, Tokyo, Japan

Overcoming the Data Delivery Bottleneck in Supercomputing

Peter Baumann

Active Knowledge GmbH
Kirchenstr. 88, D-81675 Munich, Germany
Dial-up: voice +49-89-458677-30, fax -39, mobile +49-173-5644078
E-mail: baumann@active-knowledge.de

1 Introduction

One of the well-known bottlenecks identified in climate research are the high I/O requirements of some scientific codes; other disciplines like astronomy, engineering, biology and materials science are following closely. Find below a number of examples of typical output sizes of simulations in various scientific disciplines [5].

Table 1. Typical atmospheric experiment data sizes for 10 model year run, output 4x per model day.

Spatial Resolution	Data Sizes (GB)
climate	36.5
seasonal	74.8
climatology	98.1
forecast	324.1

Table 2. Typical ocean experiment data sizes for a 10 model year run, data output once per model day

Spatial Resolution	Data Sizes (GB)
4º x 4º (global)	5.8
1º x 1º (Atlantic)	21.5
1º x 1º (global)	133.2
1/4ºx1/4º(Atlantic)	347.1

Situation becomes worse when many users, possibly concurrently, want to access result data. As a rule, user access nowadays suffers from three severe shortcomings.

First, access is on an inappropriate semantic level. Applications accessing data have to deal with directories, file names, and data formats instead of accessing multidimensional data in terms of, say, simulation space/time and other user-oriented terms.

Second, data access is inefficient. Data are stored according to their generation process, for example, in time slices. All access pertaining to different criteria, for example spatial coordinates, requires data-intensive extraction processes and, hence, suffers from severe performance penalties.

Third, search across a multitude of data sets is hard to support. Evaluation of search criteria usually requires networks transfer of each candidate data set to the client, implying a prohibitively immense amount of data to be shipped. Many interesting and important evaluations currently are impossible.

Finally, all the aforementioned data access efficiency problems are intensified as the user community grows, as obviously networks load grows linearly with the number of users.

In summary, a major bottleneck today is fast, user-centric access to and evaluation of these so-called *Multidimensional Discrete Data* (MDD).

In the European initiative ESTEDI[1,2], research and industry cooperate to overcome this obstacle. The basic approach is to augment the high-volume

[1] see www.estedi.org

[2] ESTEDI is funded by the European Commission under grant no. IST-11009.

data generators with a database system for management and extraction of spatio-temporal data.

In this contribution we introduce the ESTEDI project and give a brief overview of the RasDaMan database system.

2 The ESTEDI Project

The observation underlying the ESTEDI approach is that, whereas transfer of complete data sets to the client(s) is prohibitively time consuming, users actually do not always need the whole data set; in many cases they require some subset (e.g., cut-outs in space and time), or some kind of summary data (such as thumbnails or statistical evaluations), or a combination thereof. It is known that an intelligent spatio-temporal database server can drastically reduce networks traffic and client processing load, leading to increased data availability. For the end user this ultimately means improved quality of service in terms of performance and functionality.

The project is organised as follows. Under guidance of ERCOFTAC[3] (European Research Community on Flow, Turbulence and Combustion), represented by University of Surrey, a critical mass of large European HPC centres plus a CFD package vendor have performed a thorough requirements elicitation. In close cooperation with these partners and based on the requirements, the database experts of FORWISS and Active Knowledge GmbH have specified the common data management platform.

Currently ongoing prototype implementation of this platform relies on the multidimensional database system RasDaMan[4] which is being optimised towards HPC by enhancing it with intelligent mass storage handling, parallel retrieval, and further relevant technologies.

The architecture will be implemented and operated in a tentatively wide range of key HPC applications, forming a common pilot platform thoroughly evaluated by end-users:

- climate modelling by CLRC[5] and DKRZ[6];
- cosmological simulation by CINECA[7];
- flow modelling of chemical reactors by CSCS[8];
- satellite image retrieval and information extraction by DLR[9];
- simulation of the dynamics of gene expression by IHPC&DB[10];
- computational fluid dynamics (CFD) post-processing by Numeca International s.a.[11].

All development is in response to the user requirements crystalised by the User Interest Group (UIG) promoted by ERCOFTAC. Active promotion of the results, including regular meetings, is instrumental to raise awareness and take-up among industry and academia, both in Europe and beyond.

The project outcome will be twofold: (i) a fully published comprehensive specification for flexible DBMS-based retrieval on multi-Terabyte data tailored to the HPC field and (ii) an open prototype platform implementing this specification, evaluated under real-life conditions in key applications.

3 The Array DBMS RasDaMan

Usually, research on array DBMSs focuses on particular system components, such as multidimensional data storage [6] or formal data models [3,4]. RasDaMan, conversely, is a generic array DBMS, where "generic" means that functionality is not tied to some particular application area.

The conceptual model of RasDaMan centers around the notion of an n-D array (in the programming language sense) which can be of any dimension, size, and array cell type (for the C++ binding, this means that valid C++ types and structs are admissible). Based on a specifically designed array algebra [1], the RasDaMan query language, RasQL, offers array primitives embedded in the standard SQL query paradigm. The expressiveness of RasQL enables a wide range of statistical, imaging, and OLAP operations. To give a flavour of the query

[3] see http://imhefwww.epfl.ch/Imf/ERCOFTAC

[4] RasDaMan is a registered trademark of Active Knowledge GmbH.

[5] see www.clrc.ac.uk

[6] see www.dkrz.de

[7] see www.cineca.it

[8] see www.cscs.ch

[9] see www.dfd.dlr.de

[10] see www.csa.ru

[11] see www.numeca.be

language, we present a small example. From a set `ClimateModels` of 4-D climate models (dimensions x, y, z, t), all those models are retrieved where average temperature in 1,000m over ground exceeds 5° C. From the results, only the layers from ground up to 1,000m are delivered. The corresponding RasQL query reads as follows:

```
select cm[ *:*, *:*, *:1000, *:* ]
from   ClimeSimulations as cm
where  avg_cells(
          cm[ *:*, *:*, 1000, *:* ]
          ) > 5.0
```

Server-based query evaluation relies on algebraic optimisation and a specialised array storage manager [7,2]. Storage is based on the subdivision of an array object into arbitrary tiles, i.e., possibly non-aligned sub-arrays, combined with a spatial index [2]. Optionally, tiles are compressed using one of various algorithms.

In the course of ESTEDI, RasDaMan will be enhanced with intelligent mass storage handling and optimised towards HPC; among the research topics are complex imaging and statistical queries and their optimisation.

4 Conclusion

The ESTEDI initiative addresses a recognised major bottleneck in climate analysis: fast, user-centric access to and evaluation of simulation results. What is unique about ESTEDI is the combination of both HPC and database expertise from European institutions and beyond. We feel that such an interdisciplinary approach has considerable potential to overcome the current HPC data management bottleneck, hence we urge the interested reader to monitor ESTEDI project progress by frequently visiting www.estedi.org !

References

[1] P. Baumann: A Database Array Algebra for Spatio-Temporal Data and Beyond. *Proc. Next Generation Information Technology and Systems NGITS '99*, Zikhron Yaakov, Israel, 1999, pp. 76 - 93.

[2] P. Furtado, P. Baumann: Storage of Multidimensional Arrays Based on Arbitrary Tiling. Proc. ICDE '99, Sydney, Australia 1999, pp. 480-489.

[3] L. Libkin, R. Machlin, and L. Wong: A query language for multidimensional arrays: Design, implementation, and optimization techniques. Proc. ACM SIGMOD'96, Montreal, Canada, 1996, pp. 228 - 239.

[4] A.P. Marathe, K. Salem: Query Processing Techniques for Arrays. Proc. ACM SIGMOD '99, Philadelphia, USA, 1999, pp. 323-334.

[5] A. O'Neill, L. Steenman-Clark: Modelling Climate Variability on HPC Platforms, High Performance Computing, R.J.Allan, M.F.Guest, A.D.Simpson, D.S.Henty, D.A.Nicole (eds.), Plenum Publishing, London, 1998.

[6] S. Sarawagi, M. Stonebraker: Efficient Organization of Large Multidimensional Arrays. Proc. ICDE'94, Houston, USA, 1994, pp. 328-336.

[7] N. Widmann: Efficient Operation Execution on Multidimensional Array Data. PhD Thesis, Technische Universität München, 2000.

Appendix

The following images show sample HPC application data retrieved from RasDaMan databases. All visualisations have been done with rView, the RasDaMan visual frontend.

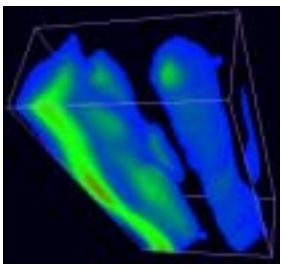

Fig. 1: Sample climate variable (3-D retrieval result from 4-D climate model); data courtesy of German Climate Research Centre (DKRZ).

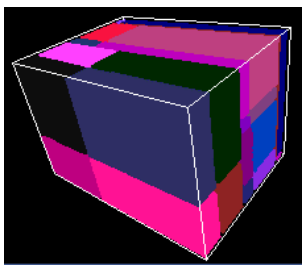

Fig. 2: Visualisation of the internal tiling structure of a 3-D array object.