The Database Research Group at ETH Zurich

Moira C. Norrie, Stephen M. Blott, Hans-Jörg Schek, Gerhard Weikum*
Computer Science Department
Swiss Federal Institute of Technology (ETH)
CH-8092 Zurich, Switzerland.
E-mail: {norrie,blott,schek,weikum}@inf.ethz.ch

1. Introduction to COSMOS

Increasingly, we are becoming a data-driven society with massive information requirements and evermore numerous on-line data sources. The research activities of the Database Group at ETH are centred on the investigation of architectures and techniques for exploring and managing the data COSMOS with its proliferation and diversity of data, and with its inherent heterogeneity. Our key aim is to provide a spectrum of data connectivity whereby data sources and application systems may cooperate at various levels of interoperability and integration. Multi-level interoperability allows application systems to cooperate with application systems, database systems to cooperate with database systems, and storage services to cooperate with storage services.

To meet this aim, we require *Cooperative Object Service Management for Open Systems* (COSMOS). The overall goal of COSMOS is to develop architectural principles for building interoperable systems that consist of cooperative data management and transaction processing services. To service the requirements of an application system as a whole, cooperation must be supported not only among components at the same level but also between components of different levels. We are therefore researching tools and techniques for both horizontal and vertical cooperation.

COSMOS bases its interoperability mechanisms on multi-level abstract object models and transaction models along with architectural principles of extensibility and openness. At the uppermost level, generic semantic object data models are used to express semantic equivalences in networks of possibly heterogeneous databases and to investigate issues of semantic interoperability. At the lower level, extensible and cooperative storage-management services provide support for diverse data management through the incorporation of specialist access methods, foreign data repositories and computational services. Sections 2 and 3 outline the research themes of the Object Data Management and Diverse Data Management Projects which deal with these upper and lower levels of interoperability, respectively. Consistency is ensured throughout via global coordination schemes and multi-level transaction management; Section 4 provides an overview of our project on Advanced Transaction Models.

Geographical Information Systems (GIS) and Computer Integrated Manufacturing (CIM) are used as driving application areas for the investigation of these issues; specific work in these areas is described in Sections 5 and 6, respectively.

Within the COSMOS, we are also investigating issues of resource management in parallel and distributed database systems. Specifically, the COMFORT project is concerned with techniques for automatic data placement in parallel storage architectures, for intelligent resource management, and for the tuning of intra-transaction parallelism in a multi-user environment. COMFORT is described in Section 7.

CONCERT is a recently initiated project to orchestrate the various COSMOS activities and provide a platform for future experimentation. The main goal of CONCERT is the vertical integration of the key concepts of the COSMOS projects to produce an advanced data management service that supports multi-level interoperability. The main features of the CONCERT system are given in Section 8.

As indicated above, COSMOS comprises a number of projects and overviews of these are given in the following sections. These projects are in part supported by a number of externally funded research programmes, some of which contribute to two or more projects. For each project its main members are listed, together with details of any external funding

*The following new affiliation of Gerhard Weikum becomes active as of April 1st, 1994: Department of Computer Science, University of Saarland, D-66041 Saarbrücken, Germany.

and a small selection of associated publications. The Database Group is large and currently consists of around twenty researchers; it also regularly hosts visiting professors. For a complete list of recent Database Group publications, please send e-mail to foerster@inf.ethz.ch.

2. Object Data Management

Moira Norrie, Michael Rys, Martin Wunderli

Object data models have been proposed not only as a basis for advanced database management systems to support complex and varied applications, but also as global models for database system interoperability. Our research efforts cover the foundations of object data models, their use for semantic interoperability in cooperative systems, and their realisation.

We are investigating the exploitation of world-wide database networks to provide global, cooperative information services based on multiple levels of integration. The level of integration among two or more nodes of a network depends on the frequency of interaction and on the level of investment in the integration effort. Our investigations span various degrees of cooperation.

At the minimum level of cooperation there is no integration between nodes and no shared schema information. A local mediator directs a query to relevant nodes and a remote node does its best to provide an answer to the query. Our approach is based on a generalisation of the ideas used in universal relational view interfaces. Nodes communicate through a minimalistic global object model which we refer to as a *universal model*. A local query is translated into a universal query which remote databases interpret in the context of their local schema.

Two externally funded projects are concerned with higher levels of cooperation and integration. In the "Databases for CIM" project, an object data model is used as a global coordination model. The coordination model must be capable of expressing global constraints and actions to be taken to ensure global consistency. Within this project, we are studying the expressive capabilities of various semantic models and the extensions necessary for modelling coordination constraints and actions. The aim of the FEMUS project is to develop a framework for federated, multi-lingual systems. In particular, the project has provided a forum for a detailed comparison of an object data model, COCOON [2], and extended Entity-Relationship Models.

The efficient realisation of object data models for the support of knowledge systems is being investigated in the context of the HYWIBAS project. Typically, knowledge systems require updates to both schema and data, and these updates may result in complex update operations at the lower levels. We are examining dynamic methods of mapping the object data model COCOON onto storage systems such that good performance is attained under various retrieval and update patterns. The exploitation of intra-transaction parallelism is a key technique towards such improved performance.

[1] M. C. Norrie, "An Extended Entity-Relationship Approach to Data Management in Object-Oriented Systems", Proc. Entity-Relationship Conf. ERA'93, Arlington, Texas, Dec 93.

[2] M. H. Scholl, C. Laasch, C. Rich, H.-J. Schek and M. Tresch, "The COCOON Object Model", Technical Report 193, Department of Computer Science, ETH Zurich, Switzerland, December 92.

Funding. HYWIBAS is funded by the Swiss National Science Foundation; the "Databases for CIM" project is funded by the KWF programme of the Swiss Federal Commission (see section 6); and FEMUS is funded by the Swiss Federal Department of Education.

3. Diverse Data Management

Stephen Blott, Helmut Kaufmann, Lukas Relly

While extensibility at the storage-management level has been extensively investigated, issues of the integration of advanced and heterogeneous application systems with those storage-management services remain vague and weakly supported. This project is concerned with the investigation of the fundamental role of database storage-management services in advanced information systems managing diverse classes of data.

A key aspect is the heterogeneity and autonomy of application systems. Such systems potentially exploit a variety of computational and indexing services in addition to pure storage-management services. This observation implies that the structure of stored data may be imposed by application systems, or by third-party services exploited by application systems, but ought not to be imposed unilaterally by the database system. It is assumed, however, that storage-management services such as persistency, transaction management and data independence remain important for the management of such data, particularly when that data is shared between multiple application systems. This project is therefore concerned with the vertical cooperation, integration and interaction between storage-management services and their application systems.

Similarly, since the data managed by advanced application systems may be highly specialised, it may be best managed by services which are themselves highly-specialised to such classes of data. For example, specialised storage and indexing services exist for the management of textual, multi-media and spatial data. Therefore, we perceive the need to provide a uniform storage-management model of application data partitioned across such horizontally-cooperating and specialised storage-management services.

A critical factor in understanding such systems is the investigation of trade-offs in the applicability of general-purpose and of specialised storage-management services. In the MUSE Project, the applicability of a relational storage-management service is being investigated for the management of textual data. A key question being addressed is that of the applicability of general-purpose technology, and the boundaries at which specialised services, say that of Information Retrieval, may better be applied. Other sub-projects include the investigation of similar issues in the management of spatial data (see Section 5), and the development of a prototype cooperative storage-management service in the context of the Concert project (see Section 8).

Funding. MUSE is partly funded by Schweizerische Rundfunk Gesellschaft (Swiss Broadcasting Corporation).

4. Advanced Transaction Models

Hans-Jörg Schek, Gerhard Weikum, Andrew Deacon, Werner Schaad, Haiyan Ye
Visitors: Prof. Yuri Breitbart (May '93–July '94), Radek Vingralek (Sept. '93–March '94)

Applications which span multiple component systems require some form of transaction mechanism to ensure global consistency. Our aim is to develop advanced transaction models which guarantee global consistency without compromising local autonomy, and which can be supported by efficient execution models. Research interests cover both the theory of transaction models and their application.

Our approach to transaction management is based on Multi-level and Open Nested Transaction Models. Access to component systems is provided through globally-accessible operational interfaces which specify locally-available operations. Operation semantics are used to determine a compatibility matrix in which potential conflicts are specified. For each operation, we define an associated compensating operation which may be used to undo actions in the event of aborts or restarts. Since our models utilise the semantics of high-level operations to increase degrees of concurrency, we refer to this as semantic transaction management.

The classical theory of transaction management is based on two different and independent criteria for the correctness of execution of transactions: the first, *serializability*, ensures the correct execution of parallel transactions in the absence of failure; the second, *strictness*, ensures correct recovery from failure. We have developed a unified model which provides a single framework for reasoning about the correctness of both concurrency control and recovery. An important advantage of our model is that it captures schedules with semantically rich operations in addition to classical read/write schedules. This therefore represents an important technique for efficient transaction processing at all architectural layers within COSMOS.

For practical studies in the application areas of CIM and banking, we are developing federated transaction management prototypes, the architectures of which follow a high-level requester-server model. Global transactions may invoke only specific, semantically-rich procedures that are exported by component systems for interoperability purposes.

A project on the use of extended transaction models for workflow management is in its early stages. Effective workflow management involves the coordination of activities across both human and computer systems in a way that is dynamic and flexible. A workflow specifies dependencies among activities and these can be represented within a common object model by means of high-level rules. To meet this challenge, we are currently investigating the integration of two database technologies – active database systems and semantic transaction management.

[1] H.-J. Schek, G. Weikum and H. Ye, "Towards a Unified Theory of Concurrency Control and Recovery", Proc. of the 12th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems (PODS), Washington DC, May 93.

[2] A. Deacon, H.-J. Schek and G. Weikum, "Semantics-Based Multilevel Transaction Management in Federated Systems", Proc. of the 10th IEEE Int. Conf. on Data Engineering, Houston, February 94.

Funding and Collaboration. The work is supported in part by the project "Databases for CIM" (see Section 6). The work on workflow management is in collaboration with Klaus Dittrich's group at the University of Zurich.

5. Data Services for GIS

Hans-Jörg Schek, Gisbert Dröge, Lukas Relly, Andreas Wolf

The overall aim of this project is to exploit database technologies to meet the demands of geographic information systems (GIS). Every geographical object contains a spatial description part. Spatial data objects require special storage services due to their wide variation in size and forms of access. In addition, spatial data requires complex processing and these operations may best be provided by external GIS specific computational services.

A specialised storage manager has been developed to support geographical objects. Spatial range queries are supported through spatial clustering which ensures efficient access to qualifying objects. A spatial access method decomposes the data space into a number of subspaces or cells which are then mapped to storage clusters. We use large multi-page storage clusters of variable size rather than single pages, and the size of a multi-page cluster is determined by a cost model. In an ideal case, an expected query is satisfied by accessing a single storage cluster of the smallest possible size; this is optimal since only exactly those objects required are retrieved. In an effort to reach this optimal performance level, the storage manager adapts the data space partition based on query ranges.

A global GIS environment may comprise multiple geo-databases and several GIS application systems. These component systems cooperate through the exchange of foreign data and/or foreign operations among the various databases and application-specific computation services. We have developed such a global GIS environment based on the extensibility of database management systems. Geographical objects can be modelled through externally defined types and the proposed mechanisms support multiple data representations and the multi-lingual working that may arise in such heterogeneous systems.

[1] G. Dröge and H.-J. Schek, "Query-Adaptive Data Space Partitioning using Variable-Size Storage Clusters", in [3].

[2] H.-J. Schek and A. Wolf, "From Extensible Databases to Interoperability between Multiple Databases and GIS Applications", in [3].

[3] *Advances in Spatial Databases*, D. Abel and B. C. Ooi, editors, Proc. 3rd Intl. Symp. on Large Spatial Databases, Singapore June 93.

Funding. This project is part of the Esprit Basic Research Action, AMUSING (Algorithms, Models, User and Service Interfaces for Geography).

6. Databases for CIM

Hans-Jörg Schek, Moira Norrie, Werner Schaad, Martin Wunderli

A CIM system supports the various activities and users involved in manufacturing systems. Such activities include computer aided design (CAD), computer aided manufacturing (CAM), production planning, parts list management and document production. We take the view that a CIM system is primarily a form of cooperative working among component systems with the emphasis on coordination rather than integration. While there is some global control to ensure system-wide consistency, existing local applications are able to operate as before and, as far as possible, system coordination is performed "behind the scenes" in such a way that global consistency is achieved with a minimum loss of autonomy.

The general aims of our work are twofold. Firstly, we are investigating ways of adapting and combining state-of-the-art multi-database technologies in real application systems and are using CIM as the driving force. Secondly, these technologies are "database technologies" and not all CIM components support database functionality. We therefore have to establish general principles for the augmentation of component systems by means of local agents such that they are enhanced with the required database functionality.

Each component system is augmented by a CIM Agent which provides a coordination interface; the interface specifies globally important local objects in terms of a global semantic object data model. System-wide consistency is ensured through the management of global constraints by a central coordinator along with multi-level transaction schemes. CIM Agents are responsible for initiating local actions necessary for the maintenance of system-wide consistency as delegated by the coordinator.

Our work is being undertaken in the context of a collaborative project involving both industrial partners and ETH's CIM-research group in the Institute of Construction and Design Methods. The project concerns include the development of enhanced CIM component systems as well as the integration of existing ones.

Funding. The project "Databases for CIM" is funded by the KWF programme of the Swiss Federal Commission.

7. COMFORT: Automatic Tuning

Gerhard Weikum, Christof Hasse, Axel Mönkeberg, Peter Zabback, Michael Rys

COMFORT stands for "Comfortable Performance Tuning". The long-term goal that we pursue in the COMFORT project is to automate, to the largest possible extent, the performance tuning of database systems. Tuning of database systems depends critically on the expertise and experience of system administrators and other human tuning experts who are responsible for the setting of system parameters. The purpose of such system parameters, or "tuning knobs", is to adapt the system to the specific characteristics of a given workload. With a wider use of OLTP and other multi-user database applications, on the one hand, and a lack of qualified tuning experts, on the other, there is a strong need for simplifying the tricky job of human administrators and ideally automating at least some critical tuning decisions.

It may appear that advances in the underlying hardware resources, such as parallel computers or very large memory, could render the need for performance tuning more or less obsolete. Parallelism may indeed greatly improve performance in some cases; however, it also bears the risk of wasting resources in other cases so that performance is possibly achieved at unacceptable cost. In a multi-user parallel database system and especially in the more realistic case of non-ideal speed-up due to skewed data, resources should be allocated carefully and tuning is crucially needed.

Our approach is to derive appropriate tuning heuristics, or "rules of thumb", from quantitative performance models for individual tuning problems, and to incorporate such heuristics in an adaptive or "self-tuning" system architecture. Thus, the goal of automatic performance tuning entails two lines of research. On the one hand, we are investigating specific tuning problems, with emphasis on the challenging problems that are posed by multi-user parallel database systems. On the other hand, we are aiming at architectural principles of a database system that can automatically adapt itself to the workload.

Within this framework of a self-tuning database system architecture, we have been addressing the following tuning issues.

1. *Data Placement in Parallel Storage Systems.* The goal is to develop algorithms and to build system software that can effectively exploit the I/O parallelism of multi-disk architectures. Since the performance of multi-disk systems depends critically on the placement of data, it is of great importance to develop algorithms for tuning the placement of data towards the workload characteristics of an application. This objective entails issues of data partitioning, data allocation, data migration and dynamic load balancing, and on-line reorganisation. We have also begun to generalise the developed methods towards distributed systems and extended storage hierarchies.

2. *Adaptive Load Control.* Load control is necessary to prevent a database system from data-contention or memory-contention thrashing, caused by excessive lock conflicts or excessive buffer replacements that may occur due to temporary load peaks. The load control method that is adopted by virtually all commercial database systems is to limit the degree of multiprogramming, that is, the maximum number of transactions that are allowed to execute concurrently. This method has the inherent limitation that it cannot react to evolving workloads. To overcome this limitation, we have been investigating adaptive load control methods that adapt the degree

of multiprogramming to the evolving transaction mix dynamically and automatically.

3. *Benefit/Cost-oriented Parallelisation of Complex Queries.* For parallelised queries with ideal speedup, response time can be improved linearly by linearly increasing the resource consumption. However, for other, less tractable queries, the marginal gain in response time that is achieved by additional resources may be so low that the additional cost would be considered as a waste of resources; in a multi-user system, these extra resources may better be assigned to other, independent queries that are invoked at the same time. Our goal is to incorporate the benefit/cost relationship of parallelised execution plans into the optimisation and processing of complex queries, and aiming at efficient, heuristic algorithms for execution plan selection and resource assignment based on benefit/cost ratio rather than response time speedup alone.

4. *Processor Allocation for Inter- and Intra-Transaction Parallelism.* In a multi-user parallel database system, resources must be divided up among many competing transactions. This can be done either by (a) servicing only few transactions in parallel, each of which is assigned many processors and other resources, or by (b) servicing many transactions in parallel, each of which is assigned only few processors and other resources. We are investigating the performance tradeoffs in the spectrum of options, aiming at intelligent heuristics for the underlying resource management problems. The goal of our approach is to adjust the degree of inter-transaction parallelism and the degrees of intra-transaction parallelism of the individual transactions to the current load dynamically and automatically.

[1] A. Mönkeberg and G. Weikum, "Performance Evaluation of an Adaptive and Robust Load Control Method for the Avoidance of Data-Contention Thrashing", Proc. of the 18th Int. Conf. on Very Large Databases, Vancouver, August 1992.

[2] P. Scheuermann, G. Weikum and P. Zabback. "Data Partitioning and Load Balancing in Parallel Disk Systems", Technical Report 209, Dept. of Computer Science, ETH Zurich, Switzerland, January 94.

[3] R. Vingralek, Y. Breitbart, G. Weikum and R. Yavatker, "Distributed File Organisation with Scalable Cost/Performance", Technical Report 207, Dept. of Computer Science, ETH Zurich, Switzerland, December 93.

Funding. The project is supported in part by the UBILAB of the Union Bank of Switzerland.

8. CONCERT: Advanced Data Service

Hans-Jörg Schek, Stephen Blott, Moira Norrie, Helmut Kaufmann, Lukas Relly, Michael Rys, Andreas Wolf, Martin Wunderli

At the upper-level, the CONCERT system provides a general-purpose collection and object manager. This provides an implementation framework for a number of higher-level and extensible object data models. It supports collection dependencies and constraints, such as those arising in the representation of generalised classification structures and entity relationships, and provides a variety of implementation strategies for managing objects and object collections.

The CONCERT Abstract-Object Kernel is the primary lower-level storage-management component of the CONCERT prototype; it may be exploited either through the collection services described above, or independently. Its design is tailored towards the management of potentially-complex application-area defined objects. A central technique towards the management of such objects is the exploitation of foreign operations to implement indexing and physical database design strategies. This requires novel approaches to aspects such as the modelling of stored objects (to support data independence), and to the importing and execution of foreign operations directly within the storage-management kernel.

The project builds on previous experiences of its participants in connection with two major projects concerned with building data management services; these are the DASDBS database kernel developed at the University of Darmstadt, and the Comandos Object Data Management Services developed as part of the Esprit Comandos project and undertaken at the University of Glasgow.

[1] H.-J. Schek, H.-B. Paul, M. H. Scholl and G. Weikum, "The DASDBS Project: Objectives, Experiences and Future Prospects", IEEE Transactions on Knowledge and Data Engineering, Vol. 2, No. 1, March 90.

[2] V. J. Cahill, R. Balter, N. Harris and X. Rousset de Pina (editors), *The Comandos Distributed Application Platform*, Springer-Verlag, 1993.