The Dresden Database Systems Group

Wolfgang Lehner
Technische Universitat Dresden – Faculty of Computer Science
01062 Dresden, Germany
wolfgang.lehner@tu-dresden.de

ABSTRACT

The *Dresden Database Systems Group* focuses on the advancement of data management techniques from a system level as well as information management perspective. With more than 15 PhD students the research group is involved in a variety of larger research projects ranging from activities to exploit modern hardware for scalable storage engines to advancing statistical methods for large-scale time series management. The group is visible at an international level as well as actively involved in cooperations with national and regional research partners.

1. INTRODUCTION

The efficient processing of large volumes of data without compromising many of the traditional database system properties like consistency, descriptive query specification, durability, etc. is one of the core pillars of many user-level applications or domains like Machine Learning. Data management solutions have therefore gained significant relevance and are also constantly faced with a wide variety of requirements ranging from application access (analytical vs. transactional) and different operator types (relational model, linear algebra, graph processing etc.) to different data characteristics (from relational rows to documents to tensors etc.). These application requirements are met by potentials and capacities on the hardware side. Especially in the recent past, hardware platforms have changed dramatically, providing substantial new opportunities for data management solutions in many areas. However, these golden prospects come along with severe constraints and increased overall system complexity.

Processing: The early multi-core era with double-digit numbers of cores per system has passed. Nowadays, multi-socket systems with up to 1,000 cores have become economically feasible. In addition, GPUs and FPGAs have made significant progress in providing general purpose processing units, but still require specific support from the software layer.

Memory: While disks are still highly usable for cold data, the increase of main memory capacities often allows to keep all working data close to the processing units. With this, the focus shifts from buffer pool management to cache optimization. In addition, non-volatile RAM will allow to directly work on primary data without copying content from the persistent to the transient memory world.

Network: Recent improvements in network technologies (e.g. Infiniband, Nx10GB Ethernet) in combination with RDMA etc. blur the boundaries between "local" (or in-node memory) and "remote" memory within a cluster, providing the opportunity to re-consider scale-up and scale-out.

Overview of research activities

Reflecting on the requirements from the application side and the opportunities on the hardware side, database systems are currently sandwiched between these two layers and have to mediate in order to provide the best service using the most efficient hardware environment. In order to holistically embrace these technological challenges and provide excellent research contributions, the *Dresden Database Systems Group* is structured into two topic areas:

System architecture: Research activities generally investigate novel system architectures as well as specific technologies to exploit modern hardware opportunities within modern storage engines. Individual topics, as detailed in Section 2, range from energy optimization via data encoding and compression to hybrid data structures for heterogeneous memory.

Data Processing: Within this field, research is conducted to push the envelope in the context of data extraction and data imputation for semi-structured data sets as well as forecasting and managing large-scale time series data. Section 3 will provide more detailed information.

Scientific environment

The *Dresden Database Systems Group* is located in Dresden (Germany), the capital of the state of Sax-

ony. Located at the heart of the Elbe valley, Dresden is famous for its baroque buildings, Mediterranean flair, and worldwide renowned cultural activities. In addition, Dresden is also one of the main research centers in Europe with important institutions like the Max Planck Society (3 institutes), Fraunhofer Society (11 institutes), Leibniz Society (4 institutes), and of course the Technische Universität Dresden (TUD), one of eleven German universities that were awarded the "University of Excellence". Moreover, Dresden is one of the largest semiconductor centers worldwide with more than 1,500 IT companies forming the region known as "Silicon Saxony".

The "Technische Universität Dresden[1]" was founded in 1828 as the "Saxon Technical School[2]" to educate workers in technological subjects such as mechanical engineering, and ship construction. Today it is among the Top-3 universities for Engineering in Germany and with approximately 37,000 students, it is one of the largest universities in Germany. The Faculty of Computer Science consists of six institutes with more than 1,700 bachelor and master students, and 180 doctoral students. Dresden has been the main research hub for computer science in Eastern Europe, making cutting-edge database research a big part of its long tradition. The database systems group is headed by Wolfgang Lehner since October 2002 and currently consists of 5 postdoctoral researchers and 15 PhD students. The group is involved in many national and international research projects and activities (Section 5). In addition to the summary below, the website of the group at `https://wwwdb.inf.tu-dresden.de/` is providing further information.

2. SYSTEM ARCHITECTURE

The group's research field in the context of efficient and scalable data processing systems embraces different research directions investigating the benefits of modern hardware and developing novel algorithms and data structures. The core question that drives the research activities is: "How should database systems be designed to optimally match new application requirements with new hardware opportunities?". To answer this question, the group develops a scalable data management platform (ERIS[3]), which is agnostic with respect to logical data models as well as physical implementations. The basic idea of this platform is to factor out as many general data management services like visibility, data and query distribution, connection management, etc. as possible and provide a plug-in mechanism for operator implementations as well as individual physical designs. In addition, the platform also systematically deploys the concept of control loops for different aspects at different levels and provides a rich set of telemetry data. For example, access statistics at the physical container level serve as input for self-optimizing access path selection. Performance counters at the CPU-level serve as input for energy optimization as well as data placement strategies. In general, this project acts as an envelope and implementation sandbox, to which individual and specific PhD projects contribute.

Energy Management

While energy consumption is a well-known issue for large-scale computing, it has also become a serious challenge in the context of individual computing systems. In this domain, our research work investigates the potentials and opportunities for fine-tuning energy consumption without compromising the overall system performance. To our own surprise, there are significant opportunities for saving energy – both from the energy efficiency and the energy proportionality perspective. Figure 1 outlines energy profiles for different workloads defined by operating individual cores and the socket infrastructure (caches, controller etc.) at different frequencies. The diagrams show individual configurations (dot size = number of active cores, dot color = average core frequency with uncore frequency in the middle) organized in a performance versus energy efficiency manner [24]. As we can see, different performance for the same work can be achieved by different configurations exhibiting different energy behaviors. With background knowledge of the type of work (column scan, hash-based aggregation, etc.) the system may pick the most energy-efficient configuration for a particular task. As we demonstrated in the context of ERIS, we can achieve up to 30% energy savings compared to the standard Linux power governor without compromising performance.

Heterogenous Systems

While using GPUs for data management activities has a long research history, most of the work focused on special implementations for highly specialized hardware configurations. In the context of heterogeneous systems research, we investigated different approaches to integrate different compute units into a single query processing environment. In [14], we propose an iteratively refined cost model to determine the optimal work distribution with respect to

[1] `https://tu-dresden.de/`
[2] `https://en.wikipedia.org/wiki/TU_Dresden`
[3] `https://wwwdb.inf.tu-dresden.de/research-projects/eris/`

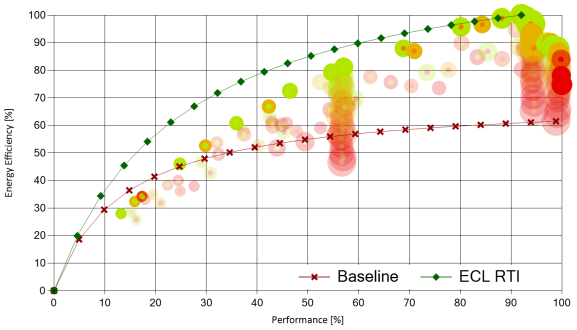

Figure 1: Energy profiles using different workloads; 12 core, 4 different core frequencies, 3 different uncore frequencies, resulting in 144 configurations

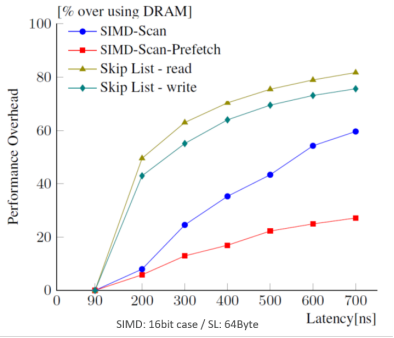

Figure 2: Performance overhead for varying memory latencies

traditional CPUs or an alternative implementation using a GPU. We also showed that the distribution of a single operations over heterogenous devices is often not practical [13] and devised an allocation scheme on a per-operator basis. The approach nicknamed HERO ("HEterogeneous Resource Optimization") provides a pseudo OpenCL device which on the one side can be registered at any OpenCL-enabled database engine and on the other side may decide on the optimal operator as well as data placement.

Non-Volatile RAM

While heterogeneity is a quite well-understood fact at the level of processing units, we see a similar trend at the level of memory. The "black-and-white"-model of RAM with buffer pool against a disk is long gone. Large main memories with different characteristics have taken over, ranging from extremely fast MCDRAM-like memories to non-volatile but still byte addressable memory systems. Within different research activities, we investigate basic characteristics and the impact on data structure design of future NVRAM. For example, Figure 2 shows that the HW-prefetcher of a CPU is able to hide most of the penalty derived from higher memory latencies for scan-based memory access patterns. However, for data structures required to follow pointers (e.g. SkipLists), the latency is directly visible as additional overhead [22].

Based on these characteristics, we developed the FP-Tree [21], a hybrid data structure spanning volatile DRAM (for the inner nodes) as well as NVRAM holding the leaf nodes for the raw data. This allows the data structure to be completely self-constrained, i.e. it does not rely on a global log but provides a micro-logging approach to bring the data structure into a consistent state after failure recovery [23]. Moreover, since HTM is a scalable method for DRAM-based data manipulation, it is inherently incompatible with NVRAM-based data structure modifications. Therefore, the FP-tree intertwines different concurrency schemes, the volatile as well as the non-volatile part. Research on efficient data structure design has a tradition within the group. A team of PhD students won the SIGMOD 2011 Programming Contest with a solution based on in-memory optimized prefix trees [17], which again was followed by the KISS-tree, a highly optimized prefix tree for supporting 32bit key lookups with exactly 3 memory accesses, independent of data cardinality and skew [18].

Data encoding schemes

The traditional disk-based layout uses a row-based or columnar data layout to represent the raw data (in combination with secondary index structures). Due to extremely high access latencies, the physical data layout of logical entities was not in the focus of optimization. Main-memory systems however demand and allow more sophisticated encoding schemes, especially compression schemes to limit the data transfer between memory and CPU as well as to increase cache utilization. Moreover, since raw as well as intermediate data are both located in main memory and therefore exhibit the same access characteristics, it seems beneficial to apply lightweight compression schemes also for intermediates.

Unfortunately, compression algorithms are highly dependent on the individual data characteristics and implementation details. Within [4], we reported on 39 different implementations of different compression algorithms ranging from logical schemes like RLE, Differential Coding, Dictionary Encoding to physical schemes like null suppression to eliminate leading zeroes in the binary representation. As expected, there is no single best algorithm, the decision is not trivial and depends on system environment as well as data characteristics. The experimental study however provides a solid base for an automated selection mechanism.

As counterpart to compression schemes, we also investigate the impact of encoding schemes for failure detection and failure discovery, which becomes more and more relevant with larger and denser main-memories. We look at applying AN coding schemes for column stores, which turns out to be a great solution for detecting multi-bit flips, as it results in a significantly lower probability for silent data corruption in combination with a simple arithmetic model. While previous work only used expensive division operations for decoding, AN coding allows transforming divisions into relatively cheap multiplications by using inverses.

3. DATA PROCESSING

As already mentioned, the research field of data processing addresses applications for managing and analyzing data. Research activities range from extracting structured data out of unstructured data to large-scale time series management.

Database Augmentation

In the era of Big Data, the number and variety of data sources is increasing every day. However, not all of this new data is available in well-structured databases or warehouses. Instead, heterogeneous collections of individual datasets such as data lakes are becoming more prevalent. This new wealth of data, though not integrated, has enormous potential for generating value in ad-hoc analysis processes, which are becoming more and more common with increasingly agile data management practices. However, in today's database management systems there is a lack of support for ad-hoc data integration of such heterogeneous data sources.

We therefore developed the entity augmentation system REA [7] that, given a set of entities and a large corpus of possible data sources, automatically retrieves the missing attributes. Due to the inherent uncertainty of the data sources and the matching process in general, REA produces not one but k different augmentations from which the user can choose. To this end, we developed an extended version of the Set Cover problem, called Top-k Consistent Set Covering, onto which we map our requirements.

On top of that, we built DrillBeyond [6] by integrating REA with PostgreSQL, that allows to combine structured and unstructured query processing and enables seamless SQL queries over both RDBMS and the Web of Data. Therefore, we designed a novel plan operator that encapsulates the retrieval part and allows direct integration of such systems into relational query processing. The operator is placed in a cost-based manner to create query plans, that are optimized for large invariant intermediate results which can be reused between multiple query evaluations.

Dresden Web Table Corpus (DTWC)

The Web has become a comprehensive resource not only for unstructured or semi-structured data, but also for relational data. Millions of relational tables embedded in HTML pages or published in the course of Open Data/Open Government initiatives provide extensive information on entities and their relationships from almost every domain. Researchers have recognized these Web tables as an important source of information for applications such as factual search, entity augmentation and ontology enrichment. Therefore, we extracted the Dresden Web Table Corpus[4] [5] a large corpus consisting of 125 million unique tables extracted from the July 2014 incarnation of the Common Crawl. The DWTC is used as a source of semi-structured data for our augmentation project but also triggered other research projects, e.g. in [3] we proposed a semantic normalization approach for Web tables containing multiple concepts, whereas in [1, 2] we proposed techniques to recover the meaning of columns by inferring knowledge base class labels and considering the Web table context.

DeExcelerator

Spreadsheets are one of the most successful content generation tools, used in almost every enterprise to perform data transformation, visualization, and analysis. The high degree of freedom provided by these tools results in very complex sheets, intermingling the actual data with formatting, formulas, layout artifacts, and textual metadata. To unlock the wealth of data contained in spreadsheets, a human analyst will often have to understand and transform the data manually. To overcome this cumbersome process, we proposed the DeExcelerator [8] that is able to automatically infer the structure and extract the data from these documents in a canonical form [19, 20].

Large-scale time series forecasting

Many analytical applications are based on empirically collected data sets derived from sensors that form large time series. Our research activities started by treating timeseries as first class citizens within a database system and introducing forecast operators to support predictive modeling. In a first step, we developed a solution that natively integrates time series forecasting into an existing DBMS, the Flash-

[4] `https://wwwdb.inf.tu-dresden.de/misc/dwtc/`

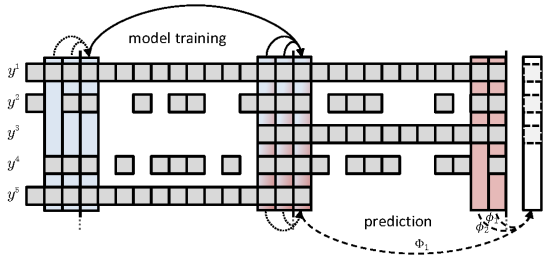

Figure 3: CSAR model with one seasonal and two non-seasonal AR components

Forward Database System (F^{2}DB) [9]. It supports a new query type–the forecast query–that enables forecasting for any database user and is transparently processed by the core engine of an existing DBMS. A key component of our system is a specialized model index structure that stores pre-built forecast models, transparently finds existing models for a given query, and maintains materialized models.

Based on this work, we reached out into multiple directions. On the one hand, we devised a novel forecasting method "The Cross-sectional Autoregression Model" (CSAR) for large-scale data sets that are highly dynamic and often noisy. While traditional forecasting approaches are focused on individual time series, resulting in a high model creation effort for a large data sets, CSAR trades depth for width by incorporating only the relevant sections of multiple time series into a model. In doing so, it provides a balance between low latency and high accuracy at individual aggregate levels. Figure 3 outlines the basic idea of CSAR with details provided–for example–in [10].

On the other hand, we investigate ways for the systematic description of time series characteristics. We developed a feature-based approach that allows us to create synthetic time series based on a given set of reference series data. As shown in [16], the approach allows users to formulate specific what-if scenarios by "tweaking" individual characteristics of the underlying time series and instantaneously see the impact in the time series data. This mechanism can be used to systematically generate time series data for simulations, model evaluation, or scalability experiments [15].

4. PARTICIPATION IN MAJOR RESEARCH ACTIVITIES

All individual research activities of the database systems group are integrated into different larger research projects funded by industrial partners, the German Research Foundation (DFG), and the European Union. The following list provides a comprehensive overview.

SAP HANA Database Campus: The group maintains a research relationship with the product development group of SAP HANA mostly located in Walldorf, Seoul, and Waterloo for more than 10 years. A variety of research activities have jointly resulted in high-profile publications as well as direct product impact. Directly involved PhD students of the Dresden group are physically located in Walldorf together with fellow PhD students from other universities forming the SAP HANA Database Campus[5].

Center for advancing electronics Dresden[6] (cfAED): cfAED was established within the German excellence initiative that represents the flagship of research funding instruments in Germany. The center aims at exploring new technologies for electronic information processing which overcomes the limits of today's predominant CMOS technology. The database systems group is actively involved in two research paths: the investigation of resilience mechanisms for data structures (using different encoding schemes), and the creation of mechanisms to bridge the gap between traditional silicon-based systems and systems based on novel materials potentially providing completely different computing characteristics.

Research Center on Highly Adaptive Energy-Efficient Computing[7] (HAEC): The HAEC project systematically and holistically investigates energy efficiency in computer systems. Starting from the hardware perspective it goes all the way up to implications for application development, compiler design, and runtime support. Wolfgang Lehner is acting a co-chairman and is responsible for all software-related activities.

Research Training Group on Role-based Software Infrastructures for continuous-context-sensitive Systems[8] (RoSI): Software with long life cycles is faced with continuously changing contexts, e.g. new functionality has to be added, new platforms have to be addressed, and existing business rules have to be adjusted. The concept of role modeling has been introduced in different fields and at different times in order to model context-related information. The central research goal of this project is to deliver proof of the capability of consistent role modeling and its practical applicability. Research activities within the database systems group try to integrate the notion of role-modeling into the database system and develop novel agile

[5] https://wiki.scn.sap.com/wiki/display/SAPHANA/Research+at+the+SAP+HANA+Database+Department
[6] https://cfaed.tu-dresden.de
[7] https://tu-dresden.de/ing/forschung/sfb912
[8] https://wwwdb.inf.tu-dresden.de/grk/

schema evolution methods to efficiently control the real-world constraints based on playing individual roles. Additionally, novel agile schema evolution methods [12, 11] are subsumed under this project [12, 11]. Wolfgang Lehner is the spokesman of this initiative, which is funded by the DFG.

Information Technologies for Business Intelligence - Doctoral College[9] (IT4BI-DC): IT4BI-DC is a doctoral program addressing six fundamental challenges in the area of Business Intelligence: Modeling and Semantics, Information Discovery, Information Integration, Business Analytics, Large-Scale Processing, and Collaboration and Privacy. The curriculum is jointly delivered by Université Libre de Bruxelles (Belgium), Aalborg Universitet (Denmark), Technische Universität Dresden (Germany), Universitat Politècnica de Catalunya (Spain), and Poznan University of Technology (Poland). Associated partners from around the world include top-ranked universities, leading industries in BI, public and private research organizations, consulting companies, and public authorities. The consortium jointly designs a set of research topics, which are jointly co-supervised by two partners of the consortium. Graduates perform their research at two of these universities and upon completion of the program are awarded with a joint degree.

5. CONTRIBUTION TO THE COMMUNITY

The Dresden Database Systems Group is supporting the database community at different levels and in different roles. At the regional and national level, Wolfgang Lehner was acting as the spokesman of the database special interest group within the "Gesellschaft für Informatik" (= German equivalent of ACM). Since April 2012, Wolfgang Lehner is elected member of the Computer science review panel of the German Research Foundation (DFG) and acts as the chairman since April 2016. At the international level, Wolfgang Lehner was member of the editorial board of the VLDB Journal from 2005 to 2011. He was Co-PC-Chair of VLDB 2011, ICDE 2015, and currently serves on the VLDB Endowment. Wolfgang Lehner is also PC member of all high-profile database conferences and was awarded with the "Distinguished PC Member" award at SIGMOD 2017. All of these activities have only been possible with a great team that is supporting and contributing. Thea team is the source of all of these fascinating research results and therefore deserves recognition for all of these remarkable achievements.

[9] `https://it4bi-dc.ulb.ac.be/`

6. REFERENCES

[1] K. Braunschweig, M. Thiele, J. Eberius, and W. Lehner. Column-specific context extraction for web tables. In *SAC*, pages 1072–1077, 2015.

[2] K. Braunschweig, M. Thiele, E. Koci, and W. Lehner. Putting web tables into context. In *KDIR*, 2016.

[3] K. Braunschweig, M. Thiele, and W. Lehner. From web tables to concepts: A semantic normalization approach. In *ER*, pages 247–260, 2015.

[4] P. Damme et al. Lightweight data compression algorithms: An experimental survey (experiments and analyses). In *EDBT*, pages 72–83, 2017.

[5] J. Eberius, K. Braunschweig, M. Hentsch, M. Thiele, A. Ahmadov, and W. Lehner. Building the dresden web table corpus: A classification approach. In *BDC*, 2015.

[6] J. Eberius, M. Thiele, K. Braunschweig, and W. Lehner. Drillbeyond: processing multi-result open world SQL queries. In *SSDBM*, pages 16:1–16:12, 2015.

[7] J. Eberius, M. Thiele, K. Braunschweig, and W. Lehner. Top-k entity augmentation using consistent set covering. In *SSDBM*, pages 8:1–8:12, 2015.

[8] J. Eberius et al. Deexcelerator: a framework for extracting relational data from partially structured documents. In *CIKM*, pages 2477–2480, 2013.

[9] U. Fischer, C. Schildt, C. Hartmann, and W. Lehner. Forecasting the data cube: A model configuration advisor for multi-dimensional data sets. In *ICDE*, 2013.

[10] C. Hartmann et al. CSAR: The cross-sectional autoregression model. In *DSAA*, 2017.

[11] K. Herrmann, H. Voigt, A. Behrend, J. Rausch, and W. Lehner. Living in parallel realities: Co-existing schema versions with a bidirectional database evolution language. In *SIGMOD*, pages 1101–1116, 2017.

[12] T. Jäkel, T. Kühn, H. Voigt, and W. Lehner. Towards a role-based contextual database. In *ADBIS*, 2016.

[13] T. Karnagel, D. Habich, and W. Lehner. Limitations of intra-operator parallelism using heterogeneous computing resources. In *ADBIS*, pages 291–305, 2016.

[14] T. Karnagel, D. Habich, and W. Lehner. Adaptive work placement for query processing on heterogeneous computing resources. *PVLDB*, 10(7):733–744, 2017.

[15] L. Kegel, M. Hahmann, and W. Lehner. Template-based time series generation with loom. In *EDBT*, 2016.

[16] L. Kegel, M. Hahmann, and W. Lehner. Generating what-if scenarios for time series data. In *SSDBM*, 2017.

[17] T. Kissinger et al. A high-throughput in-memory index, durable on flash-based SSD: insights into the winning solution of the SIGMOD programming contest 2011. *SIGMOD Record*, 41(3):44–50, 2012.

[18] T. Kissinger, B. Schlegel, D. Habich, and W. Lehner. *KISS-Tree*: smart latch-free in-memory indexing on modern architectures. In *DaMoN*, pages 16–23, 2012.

[19] E. Koci, M. Thiele, O. Romero, and W. Lehner. A machine learning approach for layout inference in spreadsheets. In *KDIR*, pages 77–88, 2016.

[20] E. Koci, M. Thiele, O. Romero, and W. Lehner. Table identification and reconstruction in spreadsheets. In *CAiSE*, pages 527–541, 2017.

[21] I. Oukid et al. Fptree: A hybrid SCM-DRAM persistent and concurrent b-tree for storage class memory. In *SIGMOD*, pages 371–386, 2016.

[22] I. Oukid and W. Lehner. Data structure engineering for byte-addressable non-volatile memory. In *SIGMOD*, pages 1759–1764, 2017.

[23] I. Oukid, W. Lehner, T. Kissinger, T. Willhalm, and P. Bumbulis. Instant recovery for main memory databases. In *CIDR*, 2015.

[24] A. Ungethüm, T. Kissinger, D. Habich, and W. Lehner. Work-energy profiles: General approach and in-memory database application. In *TPCTC*, pages 142–158, 2016.