HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Trends and Concepts in Software Industry I

Goals

Deep technical understanding of column-oriented dictionary-encoded in-memory databases and its application in enterprise computing

- Foundations of database storage techniques and operators
- Characteristics of enterprise applications and systems
- Trends in enterprise computing (e.g., machine learning)
- Hands on exercises and experiments

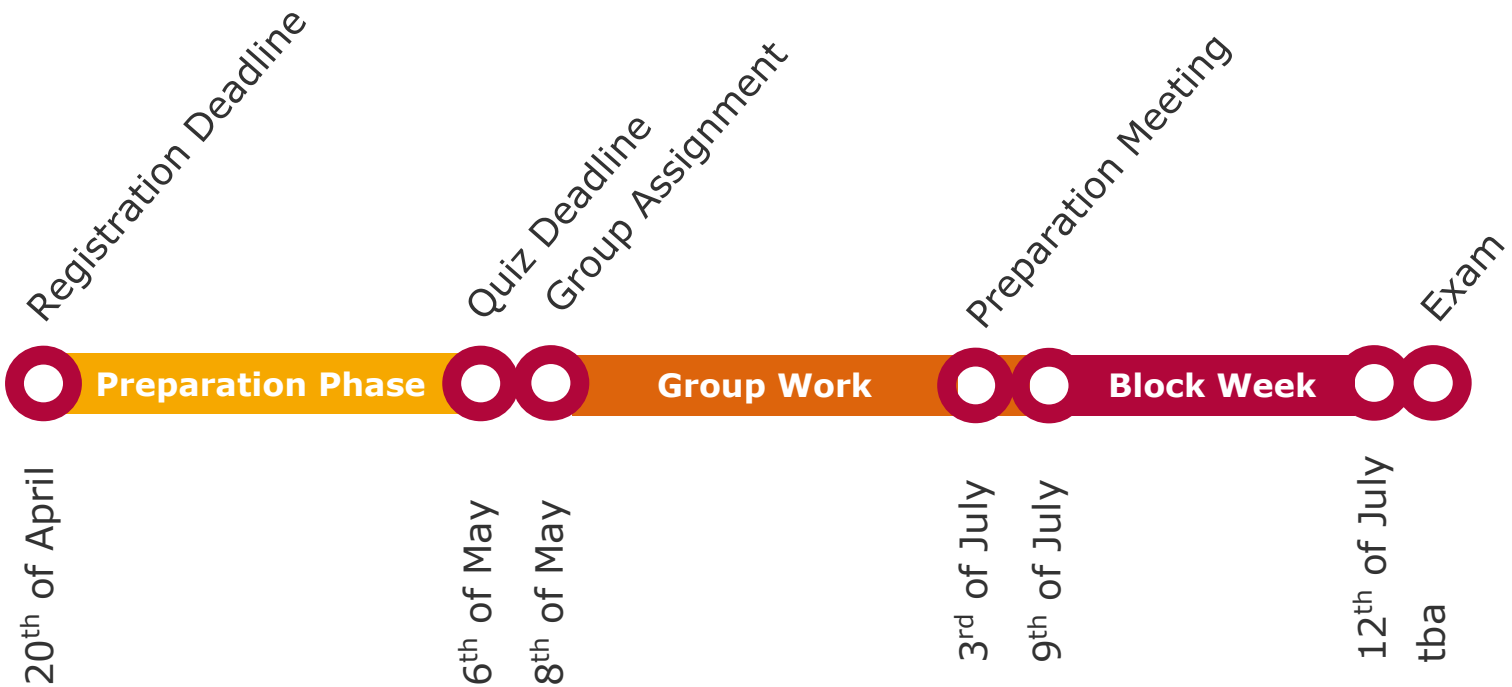
General Information

- 6 ECTS points
- Latest enrollment: 20th of April 2018
- Modules:
  - ITSE: Analyse, Entwurf, Konstruktion, and Maintenance
  - BPET: Konzepte und Methoden, Spezialisierung, and Techniken und Werkzeuge
  - ISAE: Konzepte und Methoden, Spezialisierung, and Techniken und Werkzeuge
  - SAMT: Konzepte und Methoden, Spezialisierung, and Techniken und Werkzeuge

Grading

- Final grade consists of
  - Preparation quiz (mandatory)
  - Group work, presentation, and participation during the block week (40%)
  - Written or oral exam, depending on #participants (60%)

Schedule

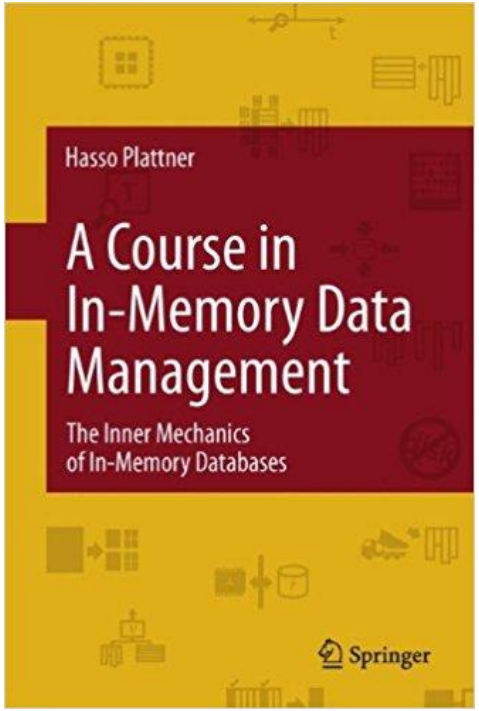

Registration Deadline
Quiz Deadline
Group Assignment
Preparation Meeting
Exam
Preparation Phase
Group Work
Block Week
20th of April
6th of May
8th of May
3rd of July
9th of July
12th of July
tba

Preparation Phase

- Get a solid understanding of the fundamentals

- Materials
  - Course book (given out by Marilena Davis, V-2.11)
  - openHPI course
    https://open.hpi.de/courses/tuk2018

- Mandatory quiz
  - Start: 23th of April
  - Deadline: 6th of May
  - Evaluation: 8th of May



Group Phase

- Preparation of interactive group part
  - Teams of 6 to 8 students guided by WiMis
  - Regular meetings
  - Team assignment: 8th of May
- Hands on experiments
  - Familiarization with existing research
  - Implementation part in C/C++
  - Evaluation of the results
  - Presentation in the block week (~30 minutes)

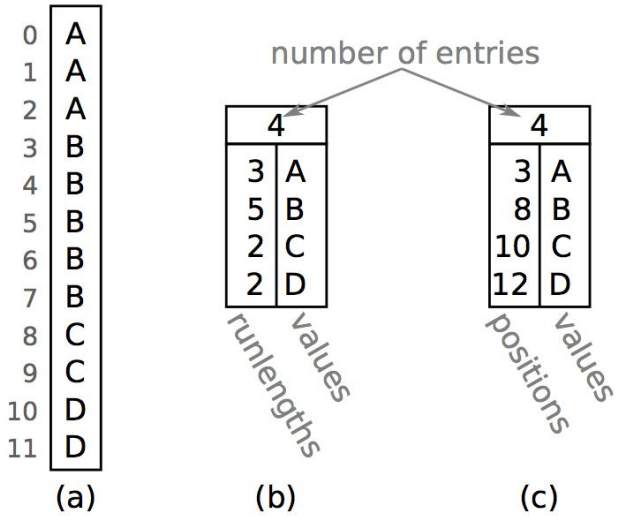
Compression techniques for column stores

Motivation

Due to latest CPU developments (many core, SIMD, NUMA) the main database bottleneck has shifted from CPU to memory bandwidth. At the same time main memory is still a critical resource in today's systems. Fortunately, compression techniques can decrease utilized bandwidth as well as memory footprint.

Experiments and Tasks

- Measure the impact of dictionary encoding and run length encoding on main memory consumption
- Measure the introduced overhead by compression/decompression (materialization)
- Compare the throughput of sequential operations on uncompressed and compressed columns



Funke et al.:Compacting Transactional Data in Hybrid OLTP&OLAP Databases

Hardware Optimization on Modern CPUs

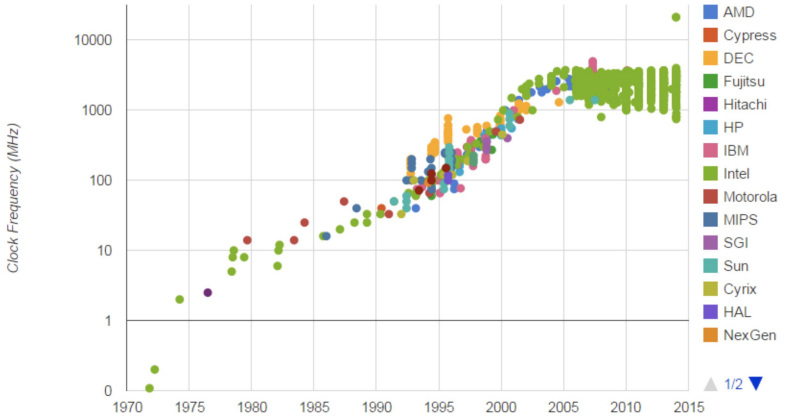
Hasso Plattner Institut

Motivation

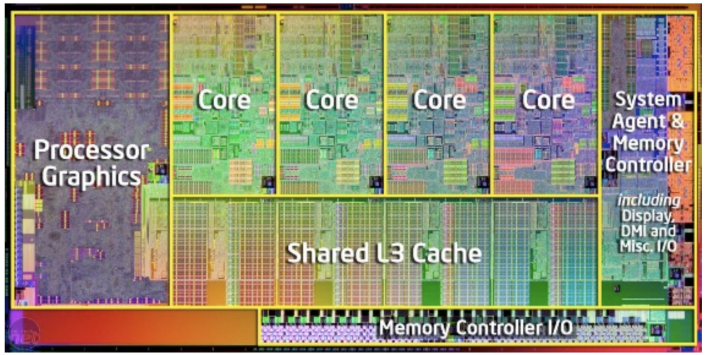
While important in the past, clock frequency has become less and less important as a measure of CPU performance. To optimize for modern CPUs, more and more factors have to be taken into account.

Experiments and Tasks

- Measure the actual main memory bandwidth for varying access patterns (random vs. sequential & single-threaded vs. multi-threaded)
- Evaluate the impact of ILP (instruction-level parallelism) on modern CPU generations and its potential pitfalls
- Measure the impact of the CPU caching hierarchy by evaluating the data locality of various access patterns on variable sized data structures



http://cpudb.stanford.edu/visualize



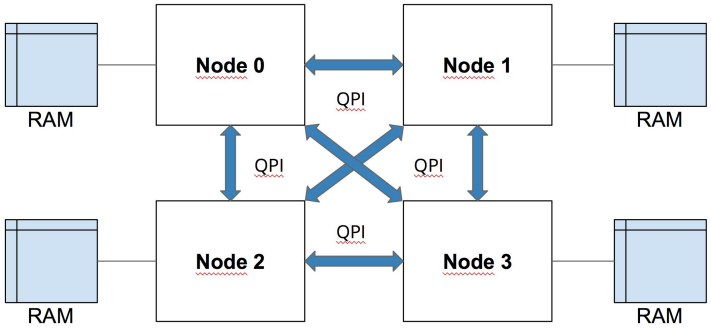
Intel

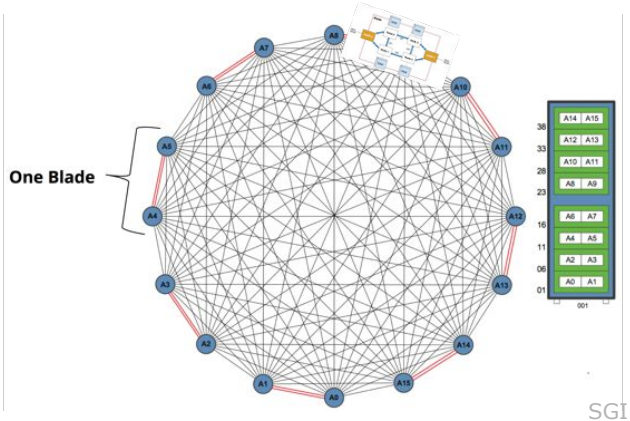
NUMA-aware optimization of data locality

Motivation
NUMA has become the standard server architecture for enterprise systems. To deliver high performance, modern databases need to be aware and leverage these hardware resources. One particular aspect of interest are memory accesses, as data can reside in local or remote memory, resulting in different costs to process data.

Experiments and Tasks

- Measure the impact of various data locations (NUMA nodes) with respect to different access patterns (sequential & random)
- Compare & measure NUMA-effects for join queries
- Which are factors influencing a query optimizer's decision in a NUMA scenario?





SGI

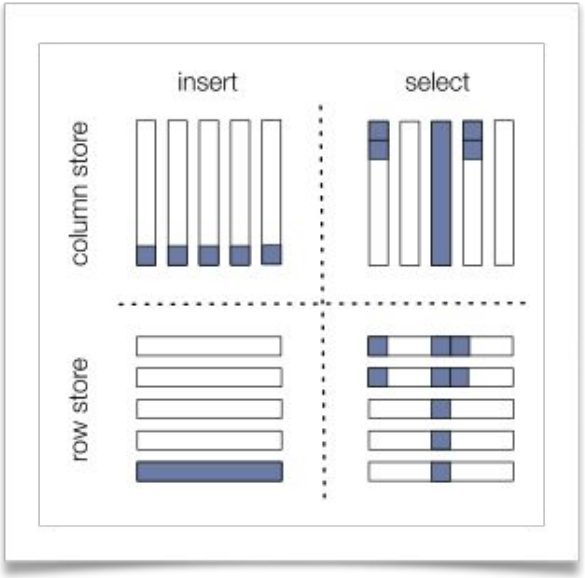
Data Layouts:
Row Orientation vs. Column Orientation

Motivation

Database systems support different data layouts. Even if the amount of data needed to process queries is the same, one or another might have advantages depending on workload and table characteristics. The reason for that is that CPU reads data in cache lines, so that sequential data access is beneficial.

Experiments and Tasks

- Measure the performance impact of #rows, selectivity, and #materialized attributes for table scans
- Measure the performance impact of #columns for inserts
- Compare the main-delta and chunk architecture wrt, scan and insert performance
- Compare the old and simplified SAP Financials data schemas with both layouts each using SAP HANA



Block Week

- General information
  - 9th of July to 12th of July
  - Lectures given by Prof. Plattner
  - Discussions about open questions in in-memory computing are a vital part of the lecture!
- Focus areas
  - Basic principles of in-memory databases
  - Characteristics of modern enterprise systems
  - Advanced data structures for in-memory databases
  - Trends in enterprise computing

Contacts

- Guenter Hesse
  - Room: V-2.03
  - Email: guenter.hesse@hpi.de
  - Phone: 1381
- Ralf Teusner
  - Room: V-2.01
  - Email: ralf.teusner@hpi.de
  - Phone: 1301

Questions