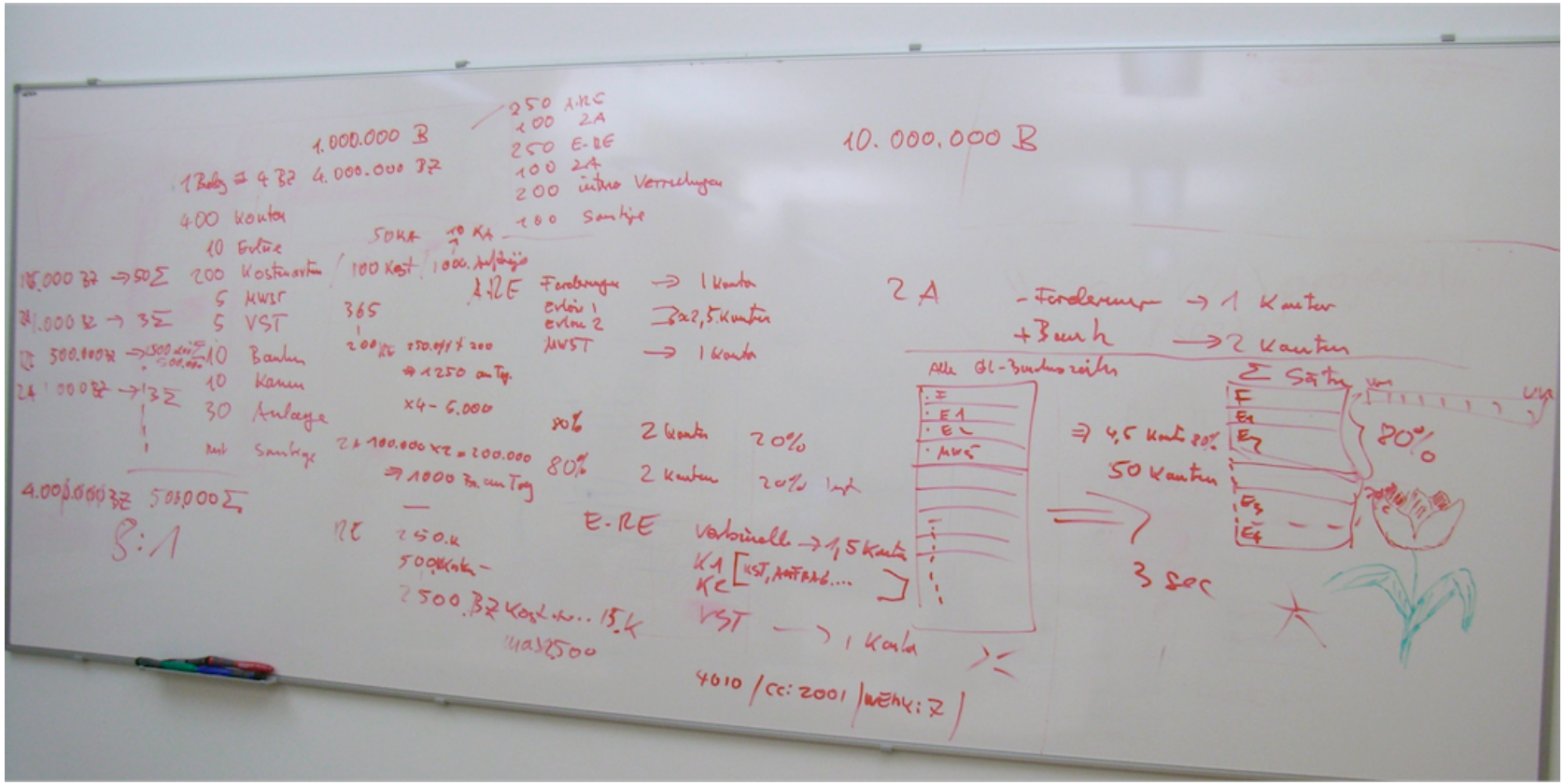
Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Trends and Concepts in Software Industry I

Columnar In-Memory Databases – The Basic Idea



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

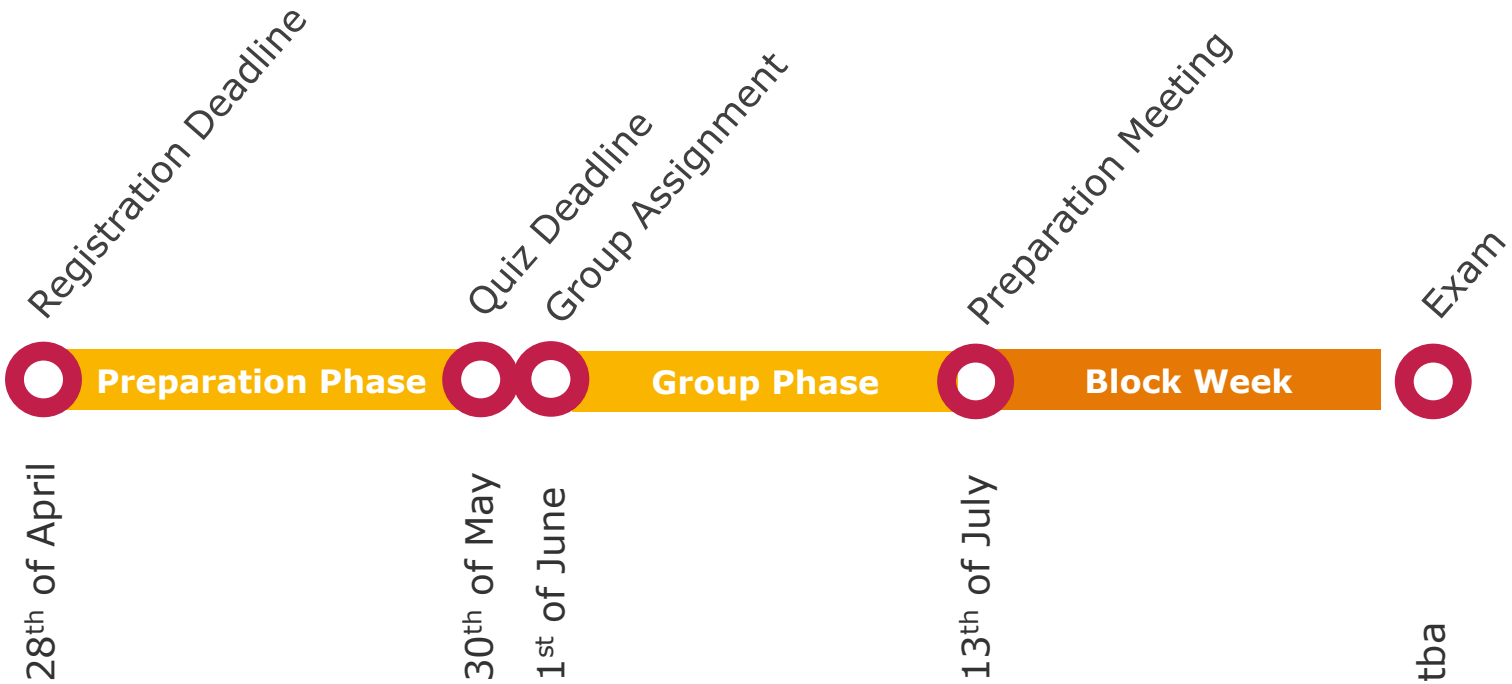
General Information

- 6 ECTS points
- Latest enrollment: 28th of April 2017
- Modules:
  - ITSE: Analyse, Entwurf, Konstruktion, and Maintenance
  - BPET: Konzepte und Methoden, Spezialisierung, and Techniken und Werkzeuge
  - OSIS: 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

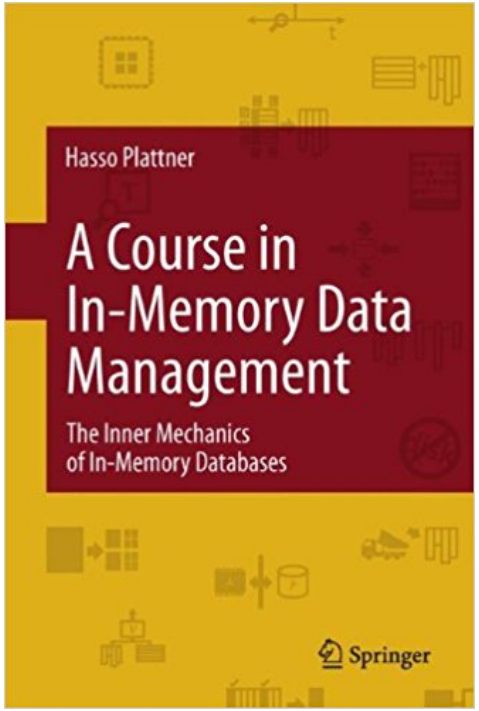

Registration Deadline
Quiz Deadline
Group Assignment
Preparation Meeting
Exam
Preparation Phase
Group Phase
Block Week
28th of April
30th of May
1st of June
13th of July
tba

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/tuk2017

- Mandatory quiz
  - Deadline: 30th of May
  - Evaluation: 1st of June



Group Phase

- Preparation of interactive group part
  - Teams of 6 to 8 students guided by WiMis
  - Regular meetings
  - Team assignment: 1st of June
- Hands on experiments
  - Familiarization with existing research
  - Implementation part in C/C++
  - Evaluation of the results
  - Presentation in the block week (~30 minutes)

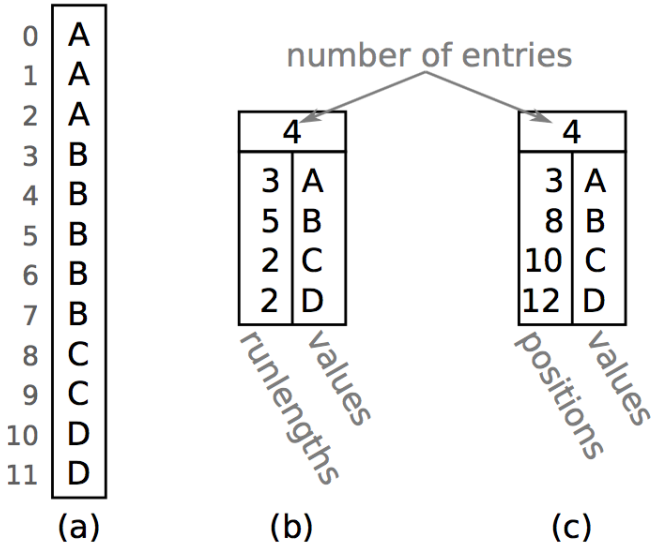
Make Databases Small Again
Compression techniques for column stores

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.

Research Questions

- What is the impact of data compression on database operators like scan and materialization?
- What is the approximate compression ratio for typical enterprise data sets?
- How much overhead is introduced by compression/ decompression?
- How do we handle frequently updated data?



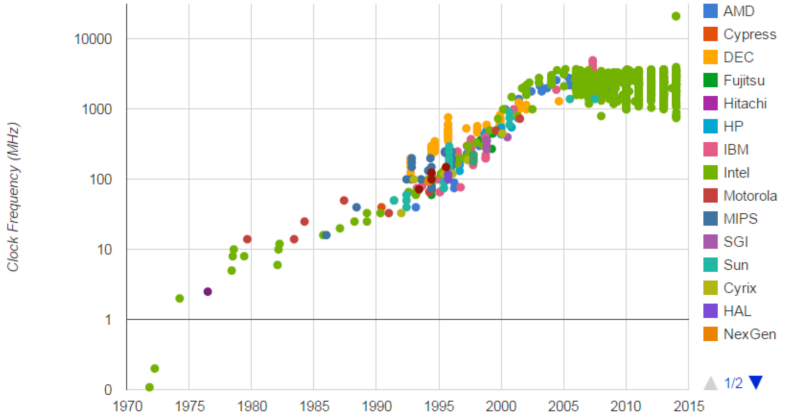
Funke et al.:Compacting Transactional Data in Hybrid OLTP&OLAP Databases

About Speed-Demons and Brainiacs
Hardware Optimizations on modern CPUs

Hasso Plattner Institut

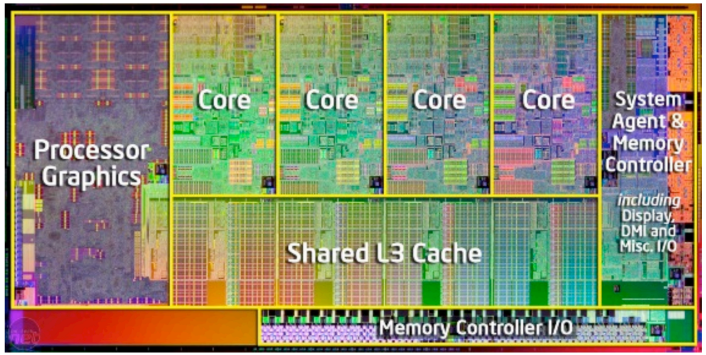
Motivation

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.



Research Questions

- How do hardware features such as Hyperthreading, Prefetching, Out-of-Order Execution, Memory Bandwidth, Execution Units, and others influence the effective performance of a CPU?
- How do they differ between CPU generations?

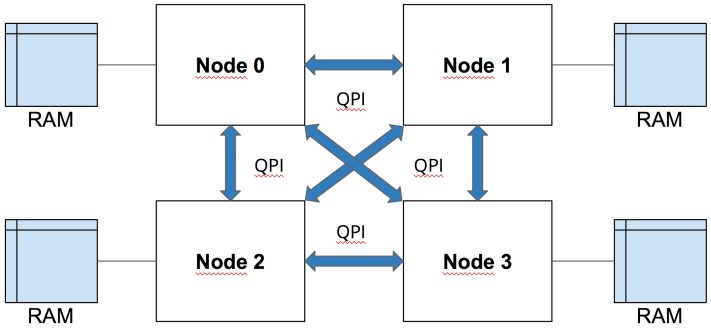


Dude: Where's my Data?
NUMA-aware optimization of data locality

Hasso Plattner Institut

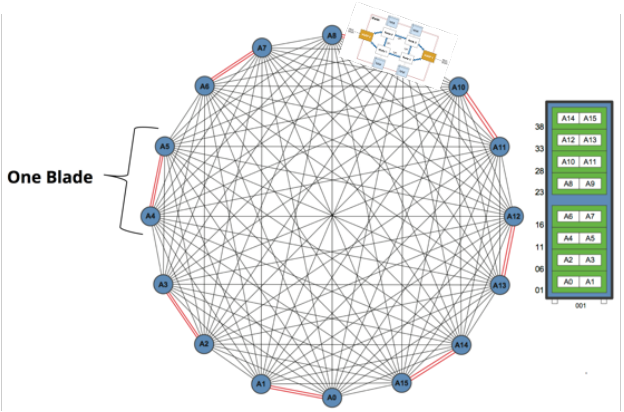
Motivation

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.



Research Questions

- What is the impact of non-local data accesses for random and sequential access patterns?
- What measure can the database undertake in order to optimize data locality?
- For cache-coherent NUMA systems: do we need to take care of NUMA at all or is the OS/CPU intelligent enough anyways?

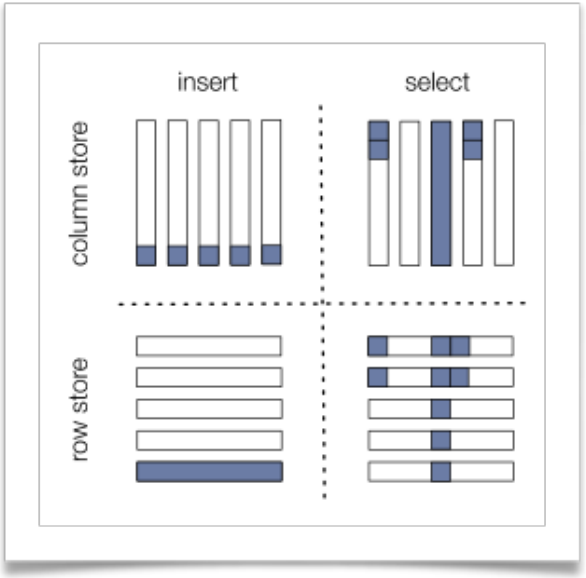


The EPIC Battle
Row Orientation vs. Column Orientation

Hasso Plattner Institut

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.



Research Questions

- How can we model insert and scan performance costs?
- What are influencing parameters WRT performance (#rows, #columns, ratio select-inserts, selectivity, #materialised attributes, etc.)?

Block Week

- General information
  - 24th of July to 27th 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

Contact Persons

- 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
- Keven Richly
  - Room: V-2.01
  - Email: keven.richly@hpi.de
  - Phone: 1328

Questions