Main-Memory Centric Data Management – Open Problems and Some Solutions

Wolfgang Lehner

Technische Universität Dresden

New Realities

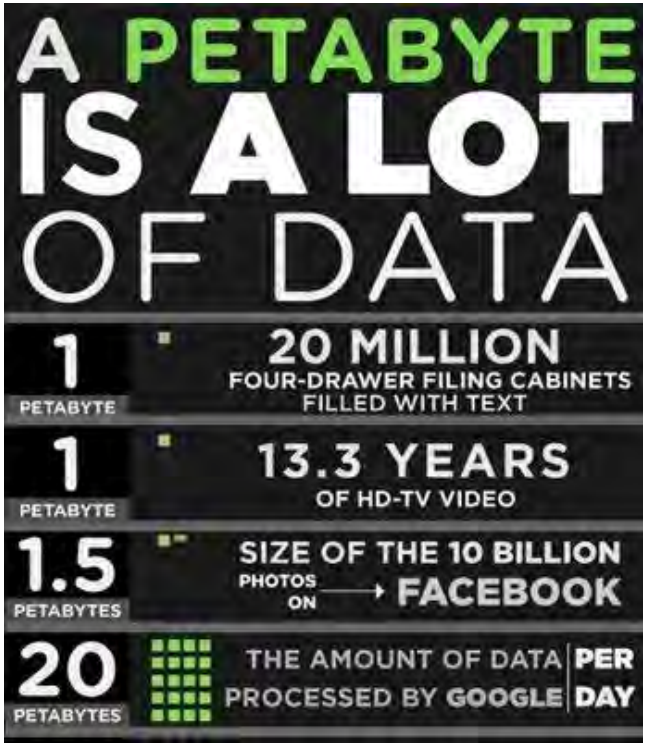
- TB disks < $100
- Everything is data
- Rise of data-driven culture
  - CERN's LHC generates 15 PB a year
  - Sloan Digital Sky Survey (200 GB/night)

The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data.

Welcome to the Petabyte Age.

The Web is a huge source of information: search engines (Google, Yahoo!) collect and store billions of documents and click streams

- 20 PB processed every day at Google (2008)
- 200 million photos are uploaded to Facebook every day → 2,314 photos/second (2010)
- 60 hours of video uploaded to YouTube every minute (2012)
- By 2015, 1 Zettabyte of data will flow over the internet per day
  (Cisco Visual Networking Index, June 2011)
- One zettabyte = stack of books from Earth to Pluto 20 times



…and considering sensors/log files/etc., every organisation is able to generate and store massive amounts of data!

TECH | 2/16/2012 @ 11:02AM | 1,530,629 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

258 comments, 149 called-out + Comment now

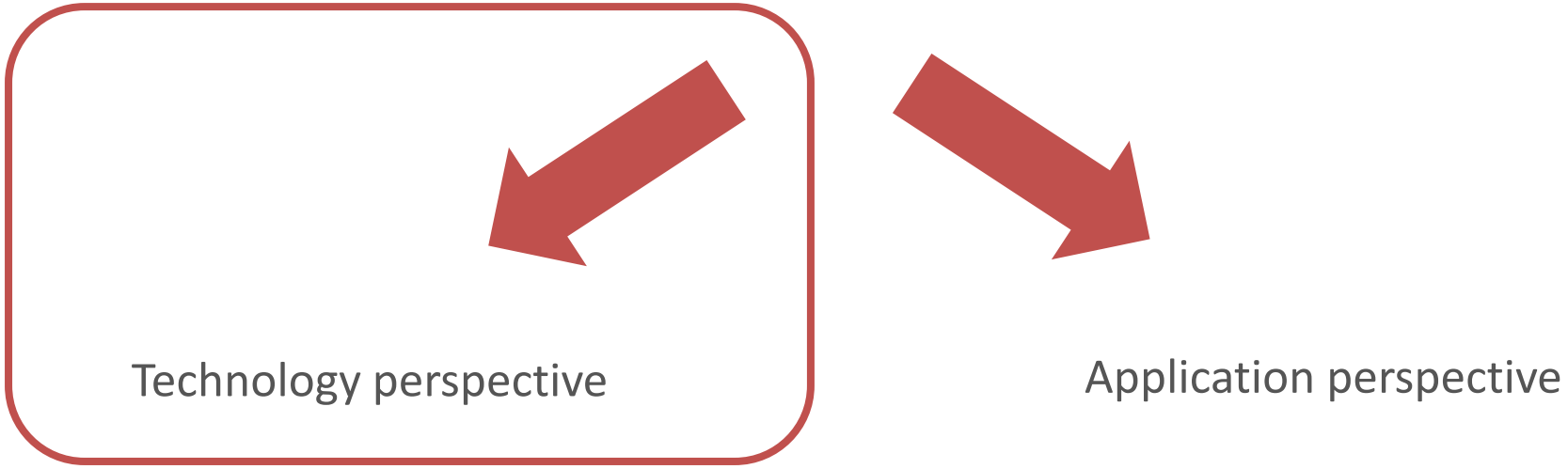
Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Target has got you in its aim

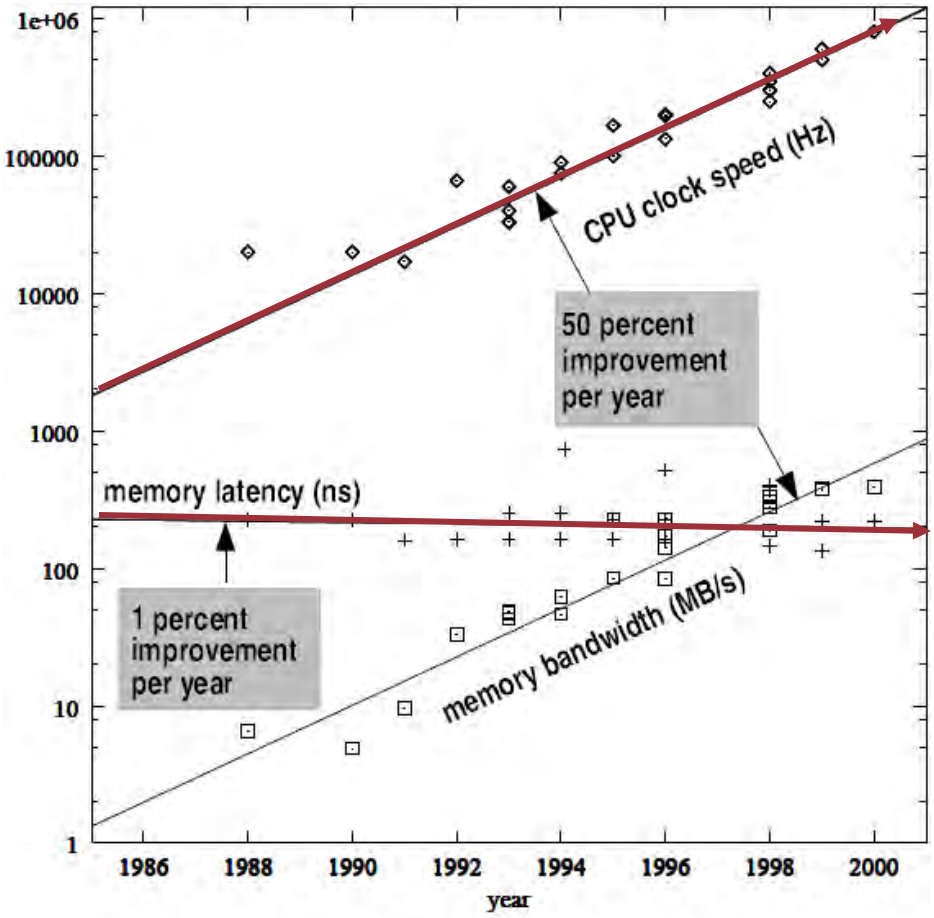
Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a

Main Memory Centric Data Management

What are the main challenges?



RAM locality is king



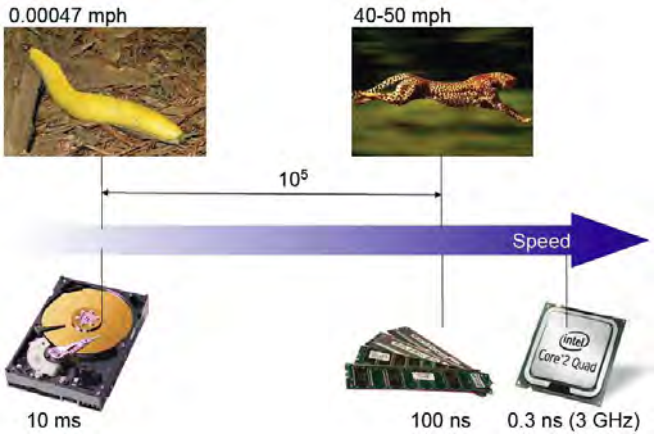
- Increased CPU calculation speed
- Increased memory bandwith
- Stagnating memory latency

- Avoid CPU idle time due to missing data
- RAM locality very important

> A Look at Current Hardware Trends

Increasing Main Memory Capacity*

„Main Memory“ is the new disk!(?)

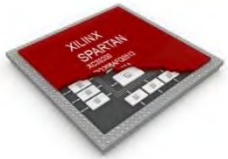


Increasing Number of Cores

- CPU/GPU, hybrids
- FPGA (Field Programmable Gate Array)

„Parallelism“ is the name of the game!

Xeon E7
20 HW Threads



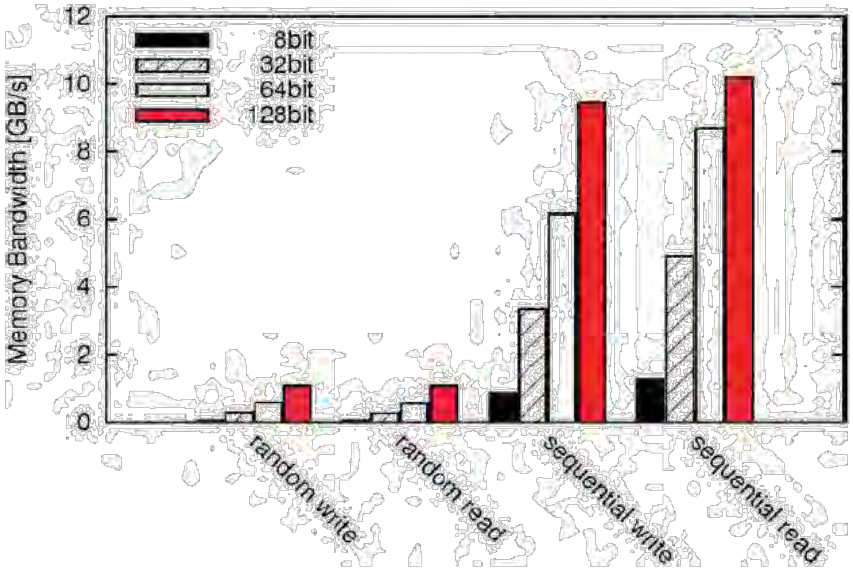
I7-2600
8 HW Threads

Knight’s Ferry
128 HW Threads

* stable RAM will be an additional ga

Faster Interconnects

> Memory Performance Comparison



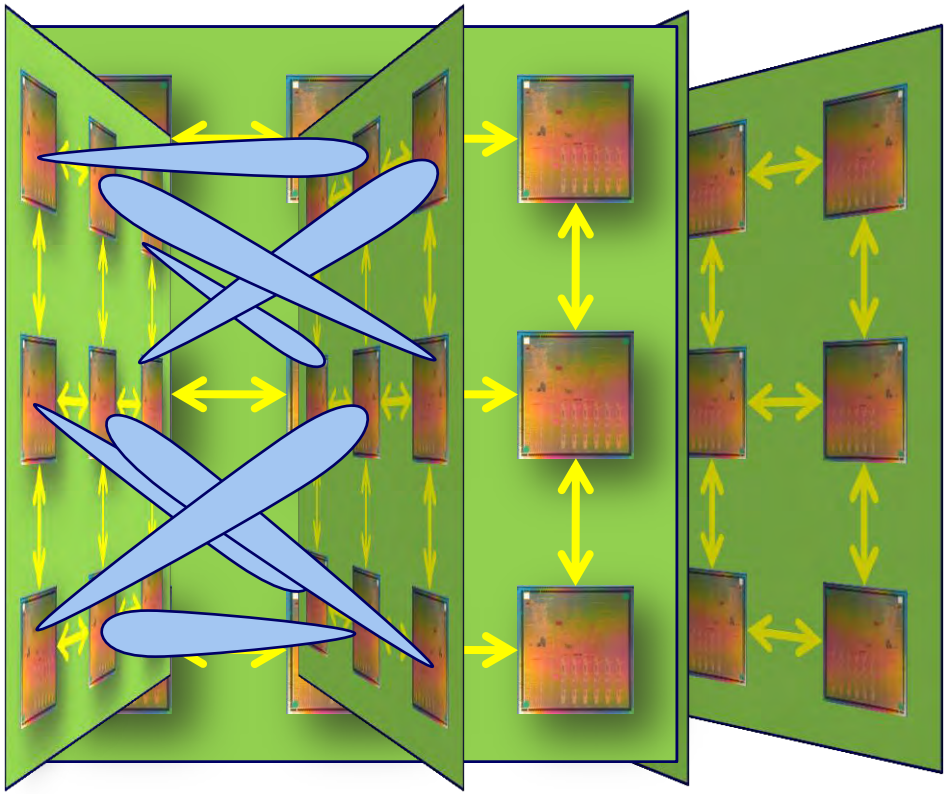
Results for a quad-core i7 2.66GHz, DDR3 1666. 32GB data accessed total.
© Tim Kaldewey

Is memory the new disk ???

- Some characteristics are very similar, e.g. random vs. sequential
- Memory architecture complicates things !

Aspect		
Rand vs. seq	1-2 orders of magnitude	3 orders of magnitude
Access granuarity	Byte addressable in theory, Caches get in the way	1 disk block, usually 4KB
Writes	Read-modify write (CL)	Read-modify-write (block)
Concurrency	Parallel memory access for peak performance	Multiple seq. streams ➔ random access

➔ Motivation for CPU-cache aware data structures, e.g. Vectorwise/X11/Ingres, SAP HANA, Greenplum, Vertica Flexstore , Oracle 11g R2 (PAX)



Optical Interconnect

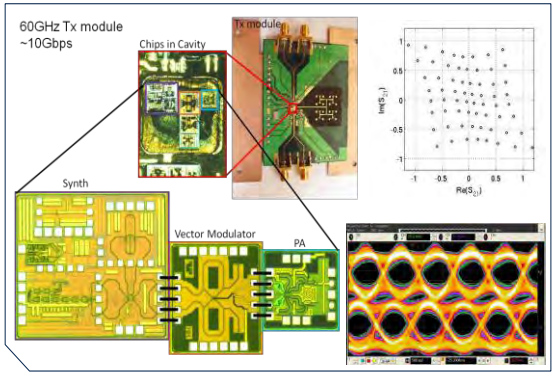
- adaptive analog/digital circuits for e/o transceiver
- embedded polymer waveguide
- packaging technologies (e.g. 3D stacking of Si/III-V hybrids)
- 90° coupling of laser

Radio Interconnect

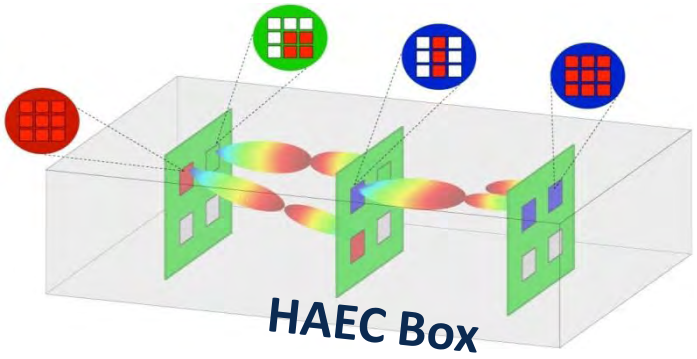
- on-chip/on-package antenna arrays
- analog/digital beamsteering and interference minimization
- 100Gb/s
- 100-300GHz channel
- 3D routing & flow management

Project A01

Millimeter-Wave Integrated Circuits for Ultra High-Speed Wireless Board-to-Board Computer Communications

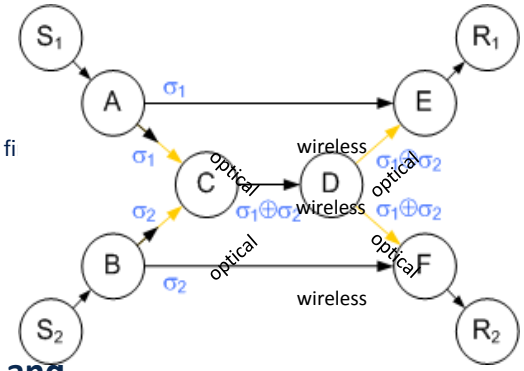


Antennas and Wave Propagation for Adaptive Wireless Backplane Communication



[SWB+10b]	BiCMOS 130 nm	122 GHz	3 GHz (2.5%)
[POZ-10]	BiCMOS 130 nm	160 GHz	~7 GHz (4%)
[BGV+09]	CMOS 65 nm	60 GHz	~6 GHz (10%)
CCN (**Tx** above)	BiCMOS 250 nm	61 GHz	10 GHz (16%)
HAEC	BiCMOS >130 nm	>150 GHz	> 20 GHz (>13%)

[GAB+10]	CMOS 65 nm	40 GS/s	18 GHz	6 bit	1.5 W	Time interl. 16 x SAR ADC
[LC10]	BiCMOS 180 nm	50 GS/s	20 GHz	5 bit	5.4 W	Time interl. 2 x flash ADC
HAEC	BiCMOS or CMOS	> 50 GS/s	> 20 GHz	1-2 bit (?)	Min.	Time interl. (?)



The fi

Sub-THz radio receiver:

Technology: BiCMOS for higher efficiency at target performance

Design approach: positive feedback for >$f_t/2$ operation

Secure Network Coding for Board-to-Board Communication

Network Coding for Wireless and Wired Onboard and Backplane Communication

reach state-of-the-art bandwidth performance for low-resolution high-efficiency ADC:

Low-resolution design: phase modulations (A02) relax amplitude accuracy, require more bandwidth

Parallelization of low-res ADC for max bandwidth

25 GHz bandwidth ADC:

Low-resolution for both energy efficiency and fastest

…a plea for specialized DB systems

The End of an Architectural Era
(It's Time for a Complete Rewrite)

Michael Stonebraker
Samuel Madden
Daniel J. Abadi
Stavros Harizopoulos

Nabil Hachem
AvantGarde Consulting, LLC
nhachem@agdba.com

Pat Helland
Microsoft Corporation
phelland@microsoft.com

Implications for the Elephants

- They are selling "one size fits all"
- Which is 30 year old legacy technology that good at nothing

Implications for the Community

Tons of research projects and startups

Self-made data management platforms

Relational stores

Document Stores

Key Value / Tuple stores

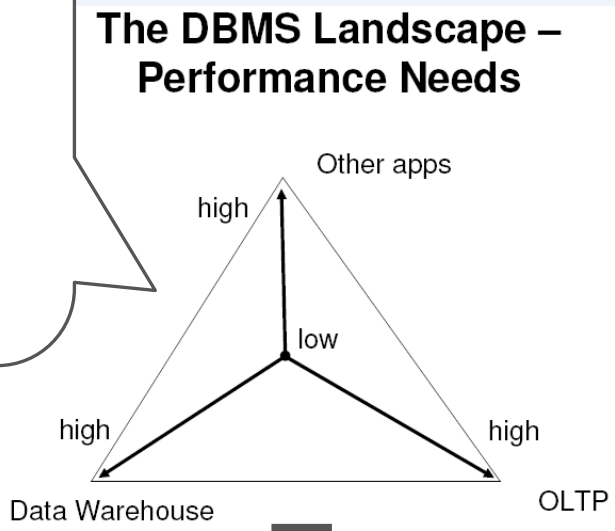
Graph Databases

Multimodel Databases

ThruDB

MemcacheDB

"Three things are important in the database world: performance, performance and performance."
Bruce Lindsey



Extreme data

Extreme performance

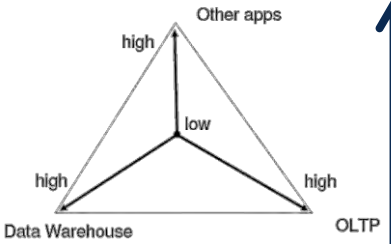
Flexibility in Database Systems

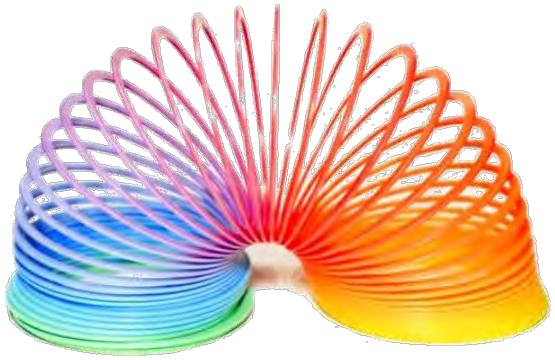
+ during deployment time
(schema definition)

- during database lifetime
(schema evolution)

- during query runtime
(scheduling, …)

Extreme Performance





Extreme Flexibility

Extreme Data

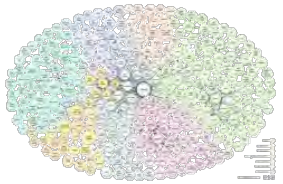
applications

data comes first, schema comes second

querying Web-Tables

role-based object models

Open Data platforms



Demand flexibility

Flexibility is cross-cutting

Database System

- *alternative query model (drill-beyond)*
- *domain-specific data models*
- *statistical operators*

data model- related

storage model- related

- record management
- logging and persistency
- HW/SW DB-CoDesign
- (MV)CC

query processing

Provide flexibility

operating system & hardware

StorageClassMemory

MegaCore systems

TeraByte-Capacity

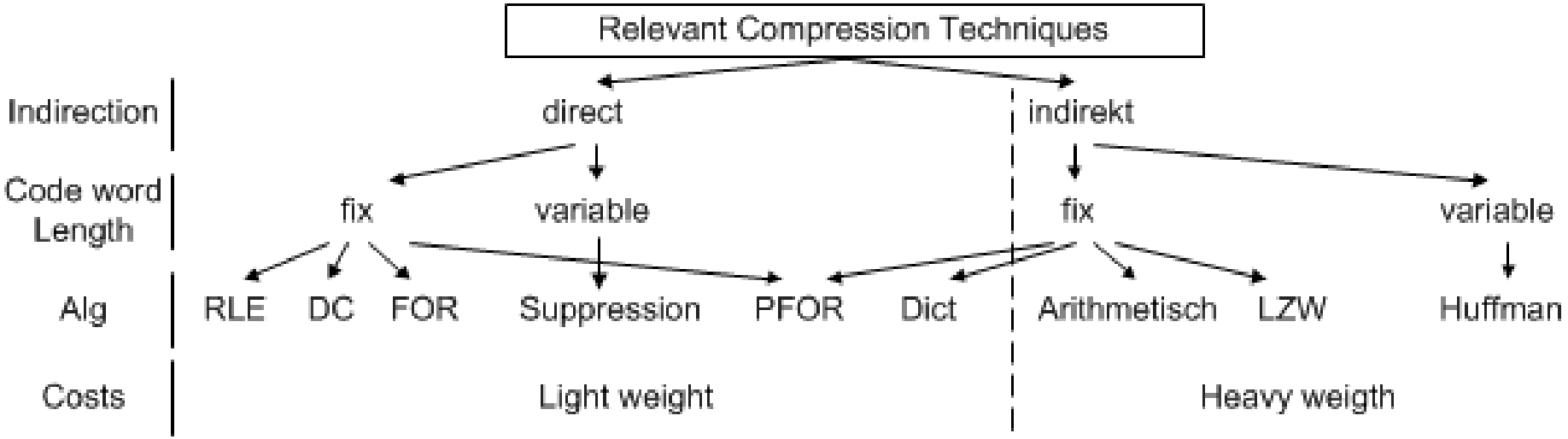
GPUs

FPGAs/FPPAs

Some Techniques

- Compression
- Indexing
- Resilience
- (Hardware) Transactional Memory

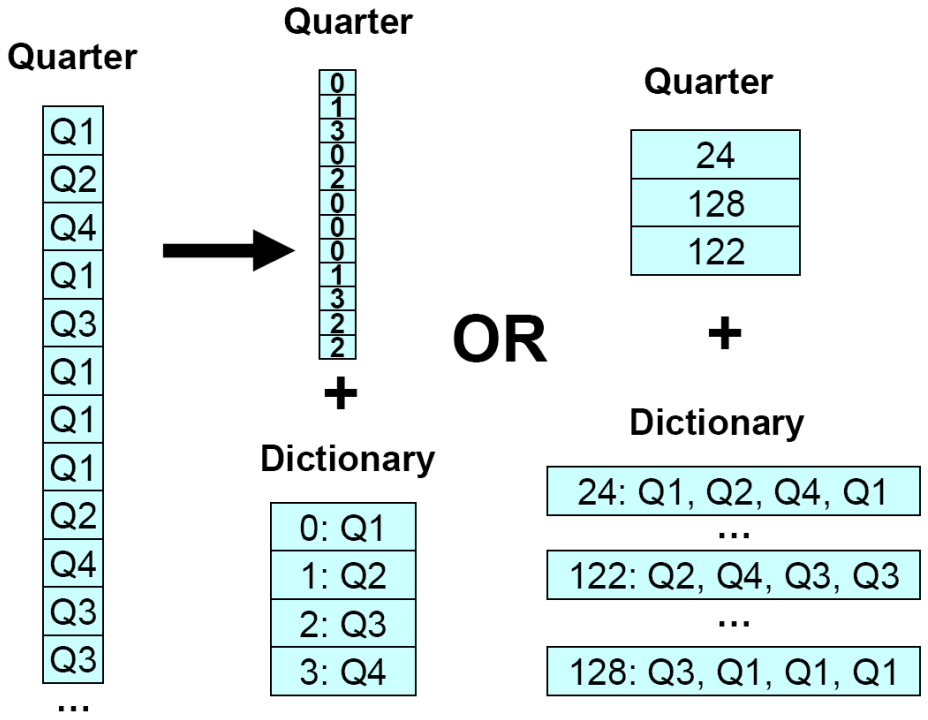
Many different compression schemes



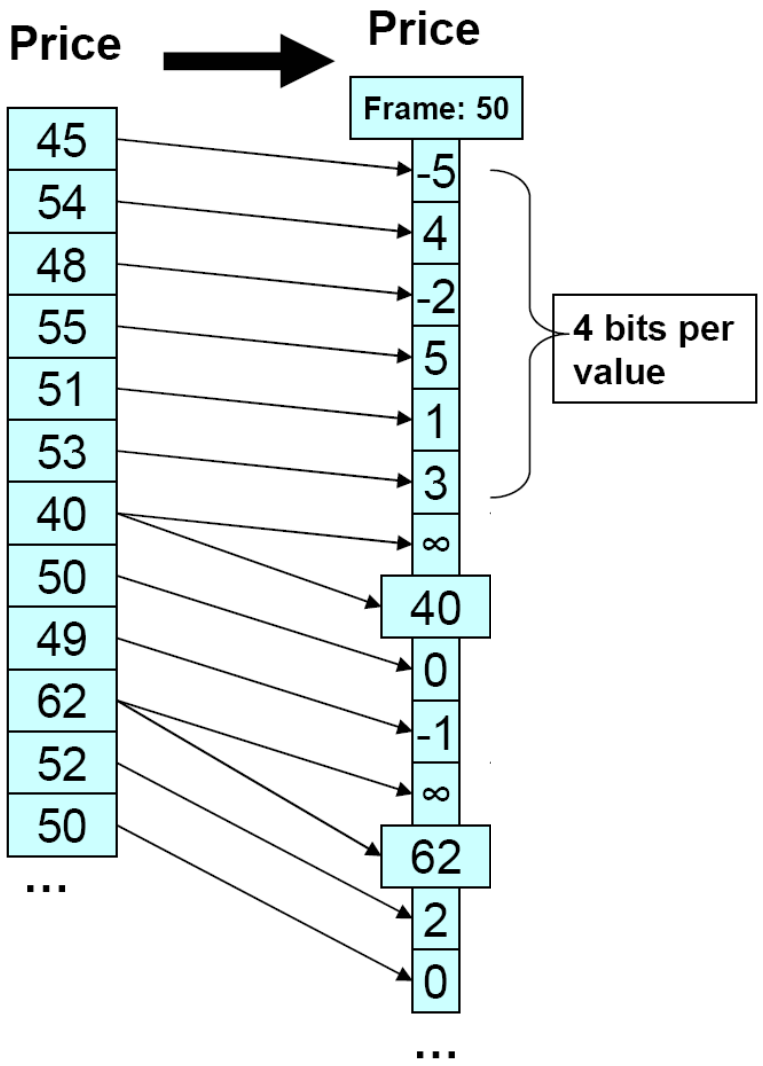
Challenges:

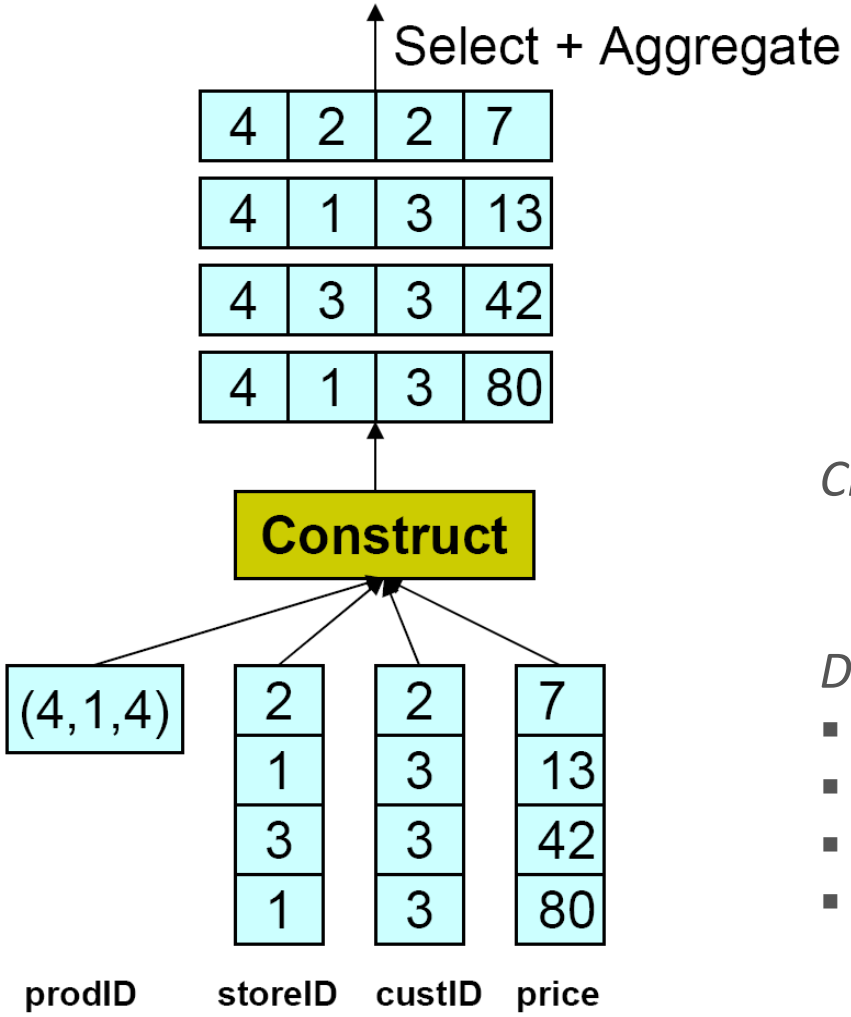
- What procedure to apply?
- When to apply compression at all?

- For each unique value create dictionary entry
- Dictionary can be per-block or per-column
- Column-stores have the advantage that dictionary entries may encode multiple values at once



- Encodes values as b bit offset from chosen frame of reference
- Special escape code (e.g. all bits set to 1) indicates a difference larger than can be stored in b bits
- After escape code, original (uncompressed) value is written





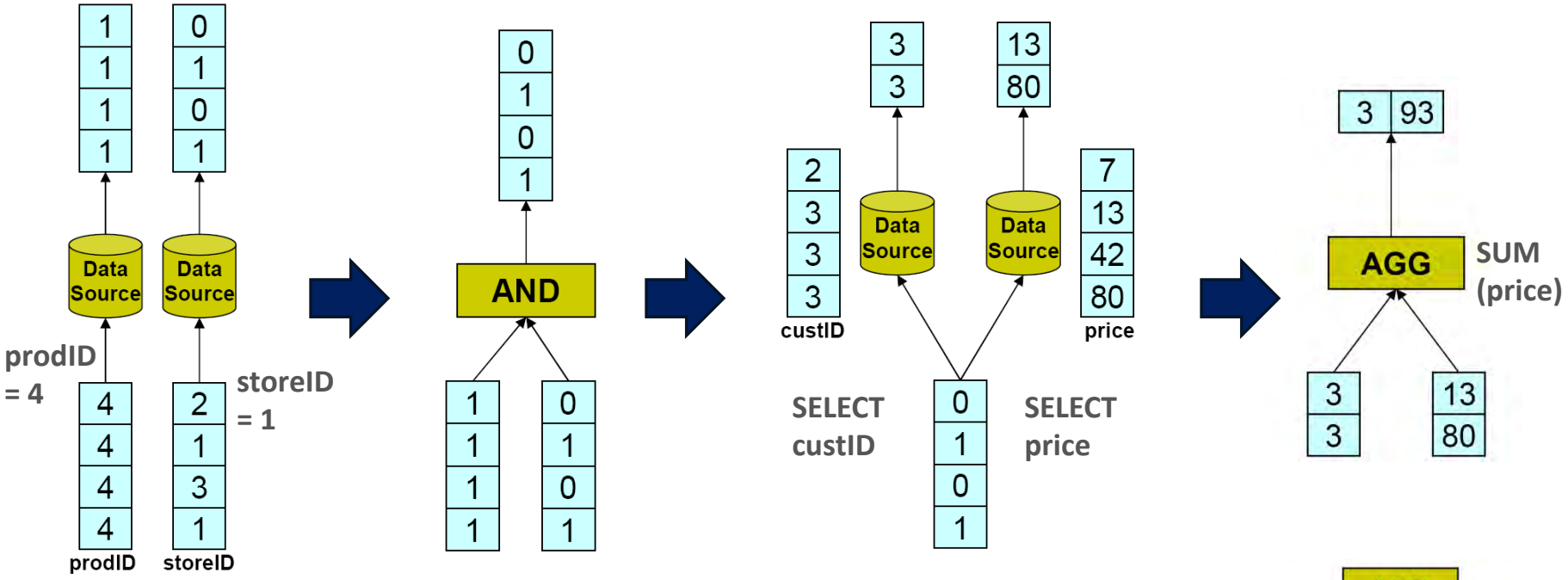

```
QUERY:

SELECT custID,SUM(price)
FROM table
WHERE (prodID = 4) AND
      (storeID = 1) AND
GROUP BY custID
```

Create rows first

Disadvantages:

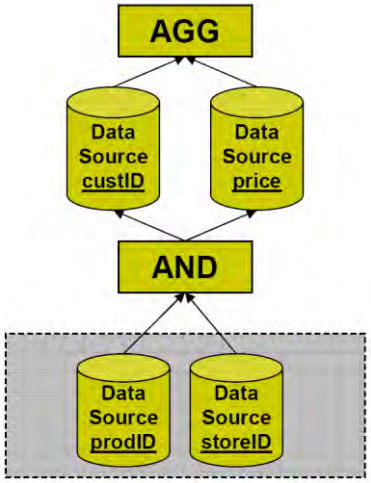
- Need to construct all tuples
- Need to decompress data
- Poor memory bandwith utilization
- Loose opportunity for vectorized operation



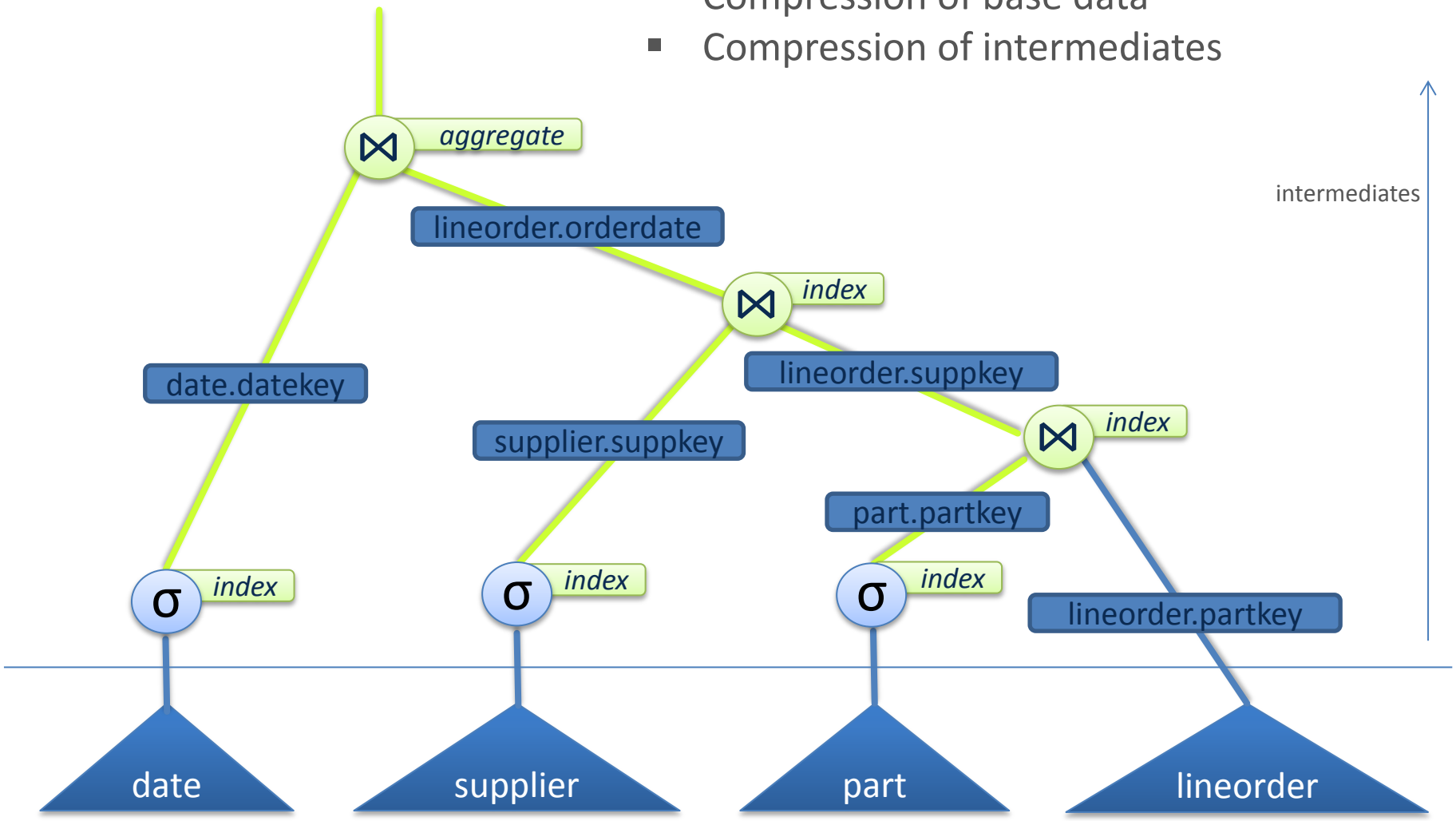

```
QUERY:

SELECT custID,SUM(price)
FROM table
WHERE (prodID = 4) AND
      (storeID = 1) AND
GROUP BY custID
```

prodID	storeID	custID	price
4	2	2	7
4	1	3	13
4	3	3	42
4	1	3	80



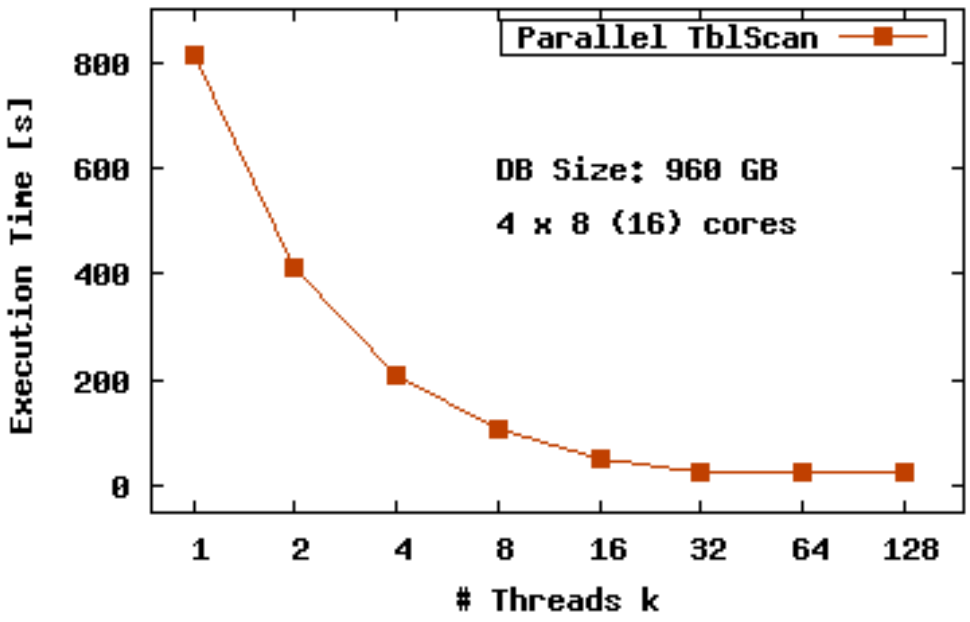
- Each operator adjusts its output to the requirements of the successive operator
  - Compression of base data
  - Compression of intermediates

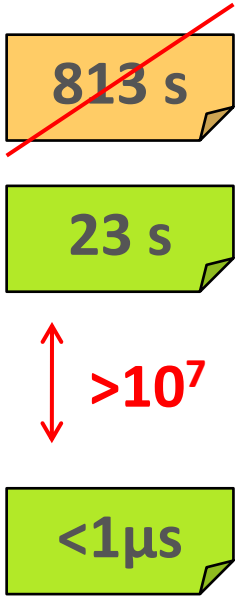


Some Techniques

- Compression
- Indexing
- Resilience
- (Hardware) Transactional Memory

- About 40 GB/s scan performance with in-memory Databases
- Real-Time Analytics requires low response time

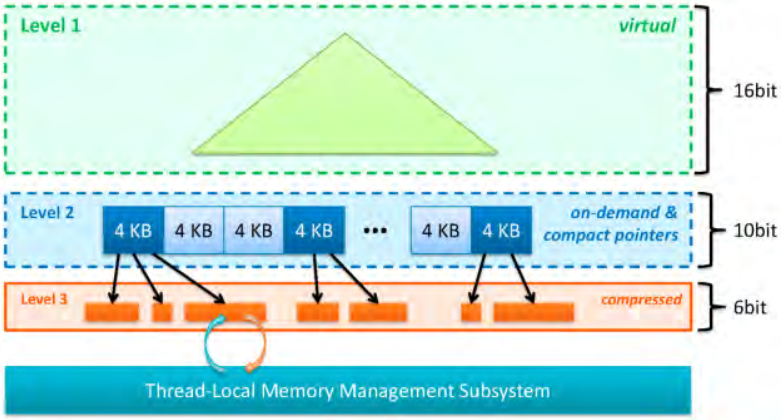




→ Indexing is still necessary for ordering, grouping, point- and range-queries

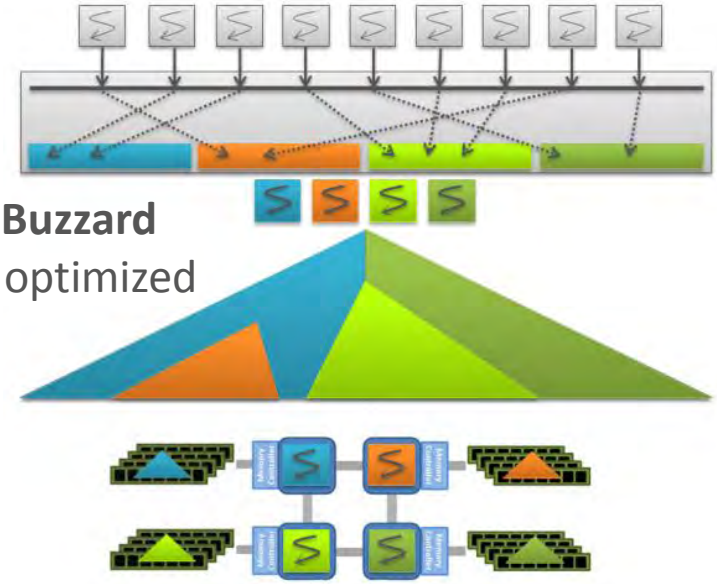
KISS-Tree

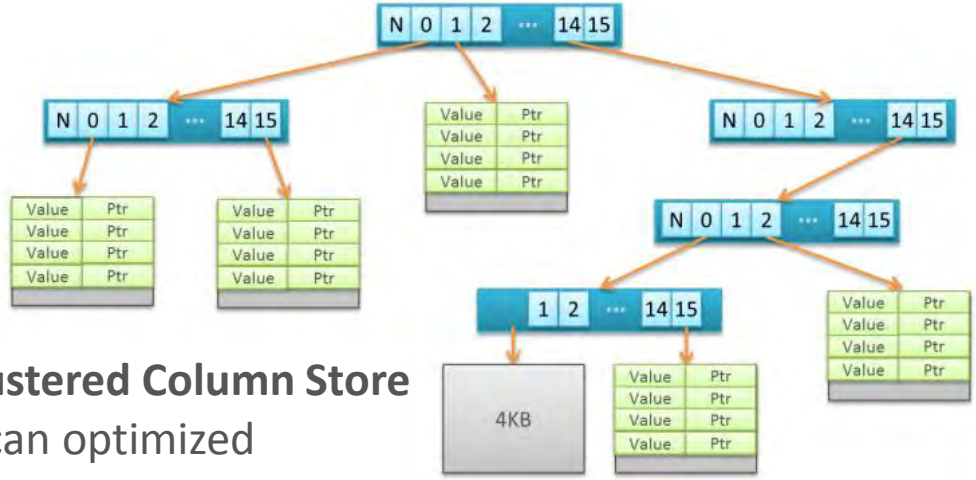
→ 32bit key index



Buzzard

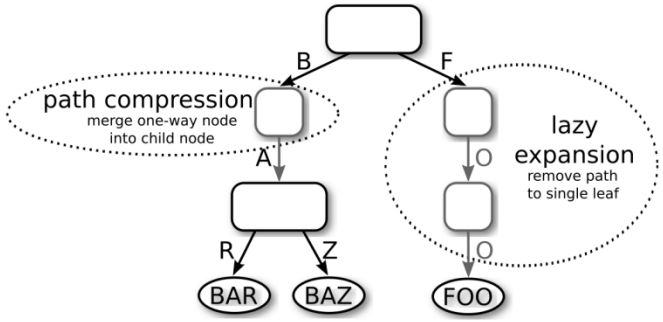
→ NUMA optimized





Clustered Column Store

→ Scan optimized



Adaptive Radix Tree (ART)

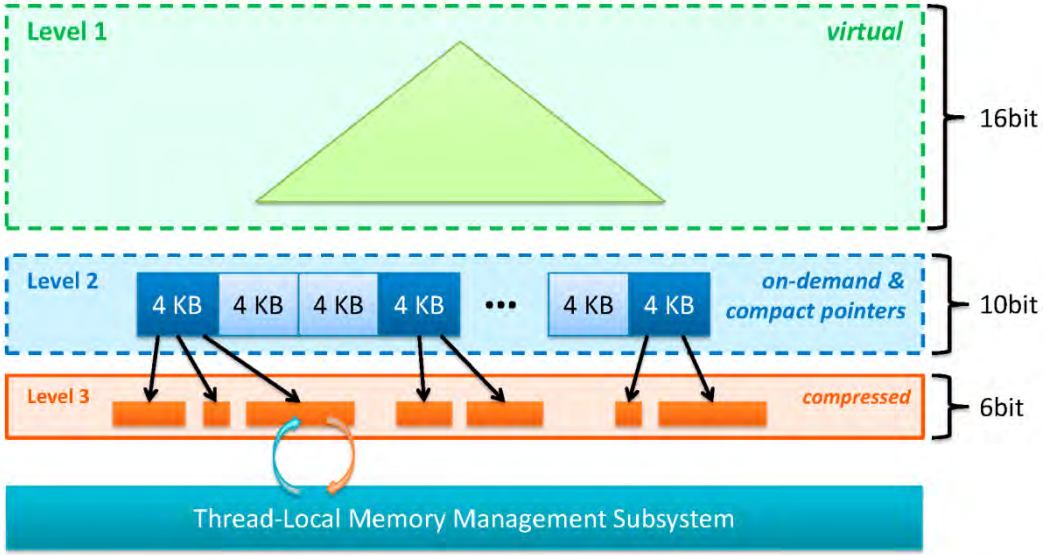
> KISS-Tree Overview

Properties

- Specialized version for 32bit keys
- **Latch-free updates**
- **Order-preserving**
- **2-3 memory accesses per key**

→ Comparable fast to reported order-preserving in-memory indexes for read access

→ BUT:
High update performance



- Heterogeneous in-memory index structure
- Combination of direct and indirect addressing
- Takes advantage of virtual memory management
- Enables different compression mechanisms

DAEMON 2012

***KISS-Tree:* Smart Latch-Free In-Memory Indexing on Modern Architectures**

Thomas Kissinger, Benjamin Schlegel, Dirk Habich, Wolfgang Lehner
Database Technology Group
Technische Universität Dresden
01062 Dresden, Germany
{firstname.lastname}@tu-dresden.de

ABSTRACT

Growing main memory capacities and an increasing number of hardware threads in modern server systems led to fundamental changes in database architectures. Most importantly, query processing is nowadays performed on data that is often completely stored in main memory. Despite of a high main memory scan performance, index structures are still important components, but they have to be designed from scratch to cope with the specific characteristics of main memory and to exploit the high degree of parallelism. Current research mainly focused on adapting block-optimized B+-Trees, but these data structures were designed for secondary memory and involve comprehensive structural maintenance for updates.

In this paper, we present the *KISS-Tree*, a latch-free in-memory index that is optimized for a minimum number of memory accesses and a high number of concurrent updates. More specifically, we aim for the same performance as modern hash-based algorithms but keeping the order-preserving nature of trees. We achieve this by using a prefix tree that incorporates virtual memory management functionality and compression schemes. In our experiments, we evaluate the *KISS-Tree* on different workloads and hardware platforms and compare the results to existing in-memory indexes. The *KISS-Tree* offers the highest reported read performance on current architectures, a balanced read/write performance, and has a low memory footprint.

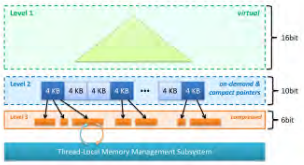
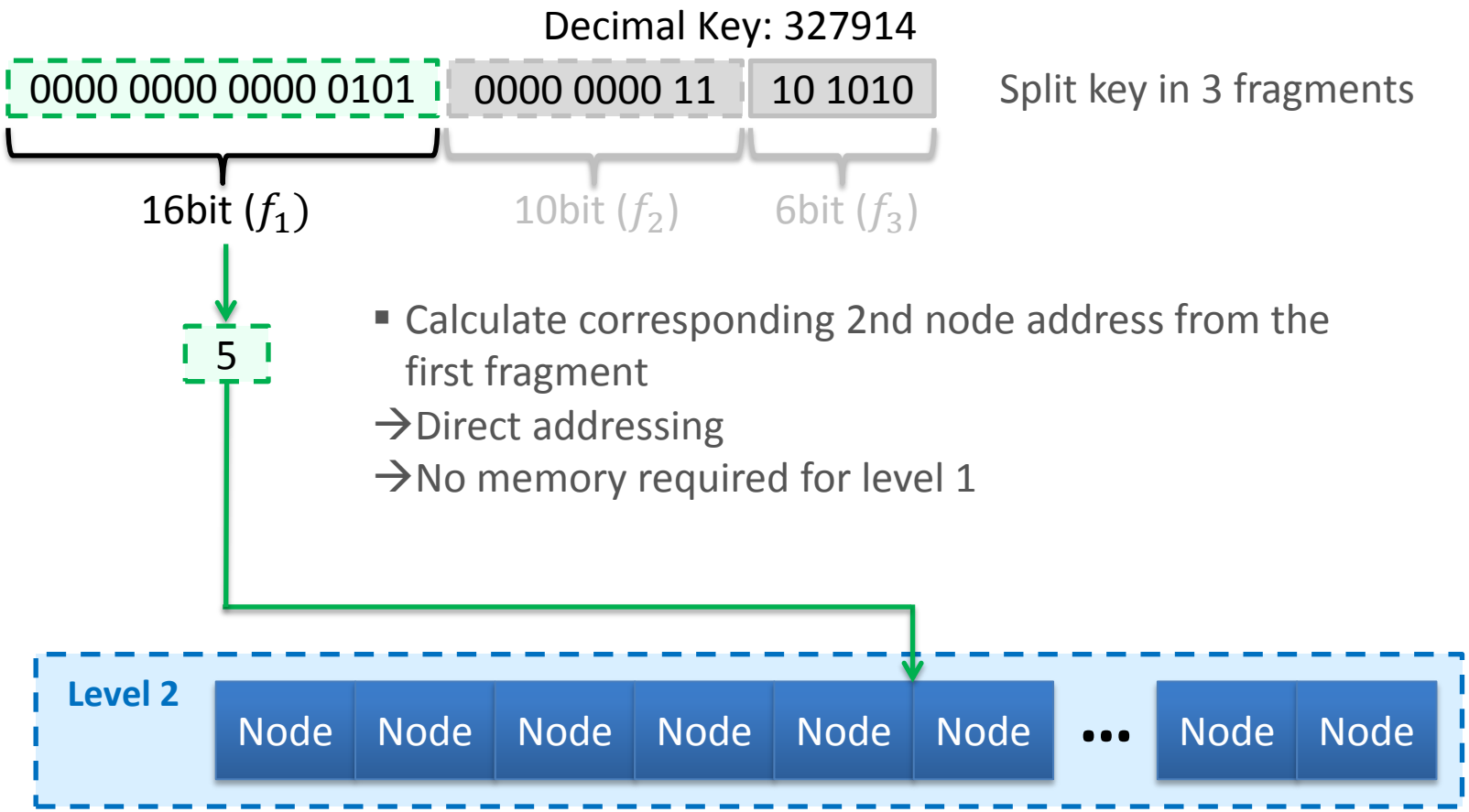


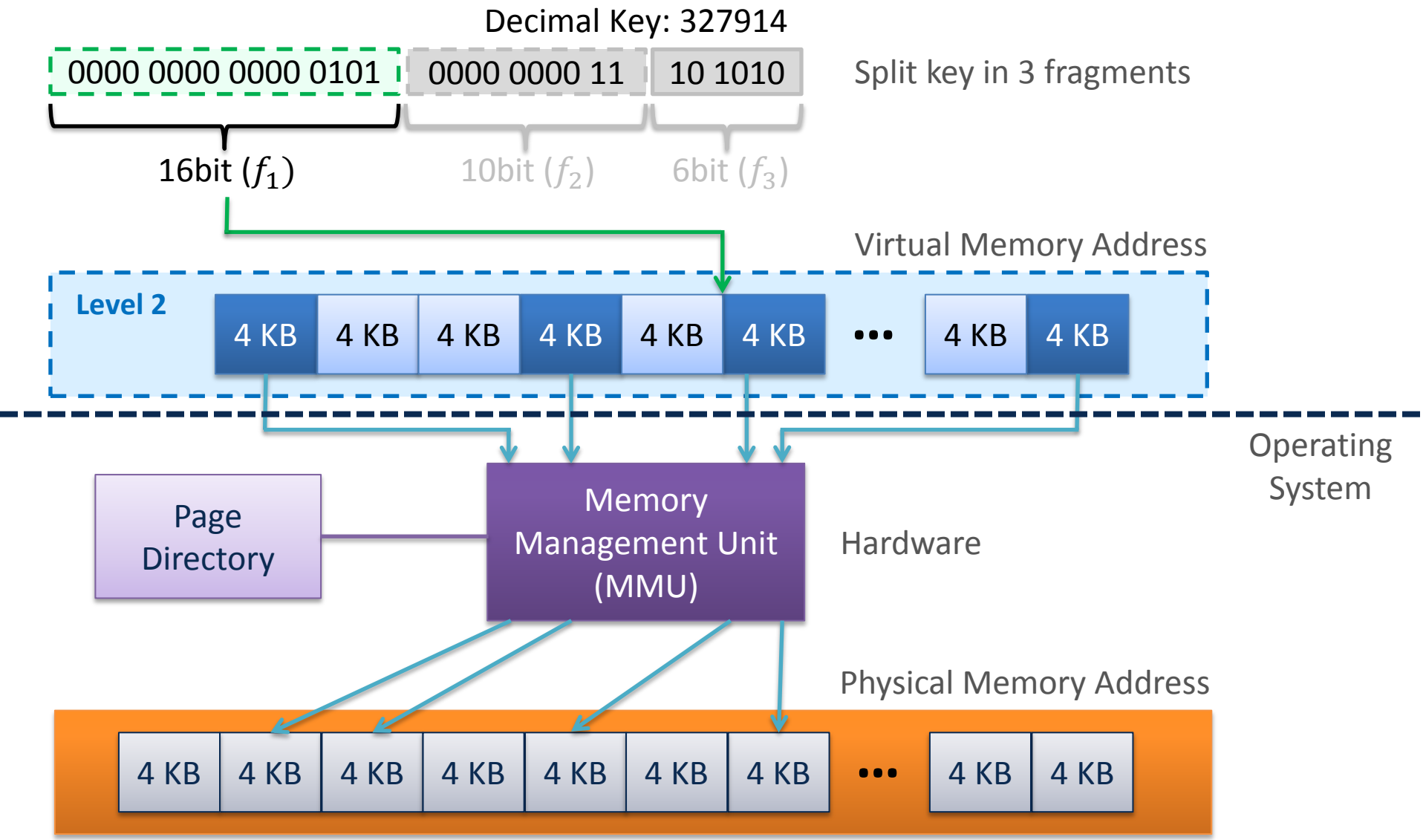
Figure 1: KISS-Tree Overview.

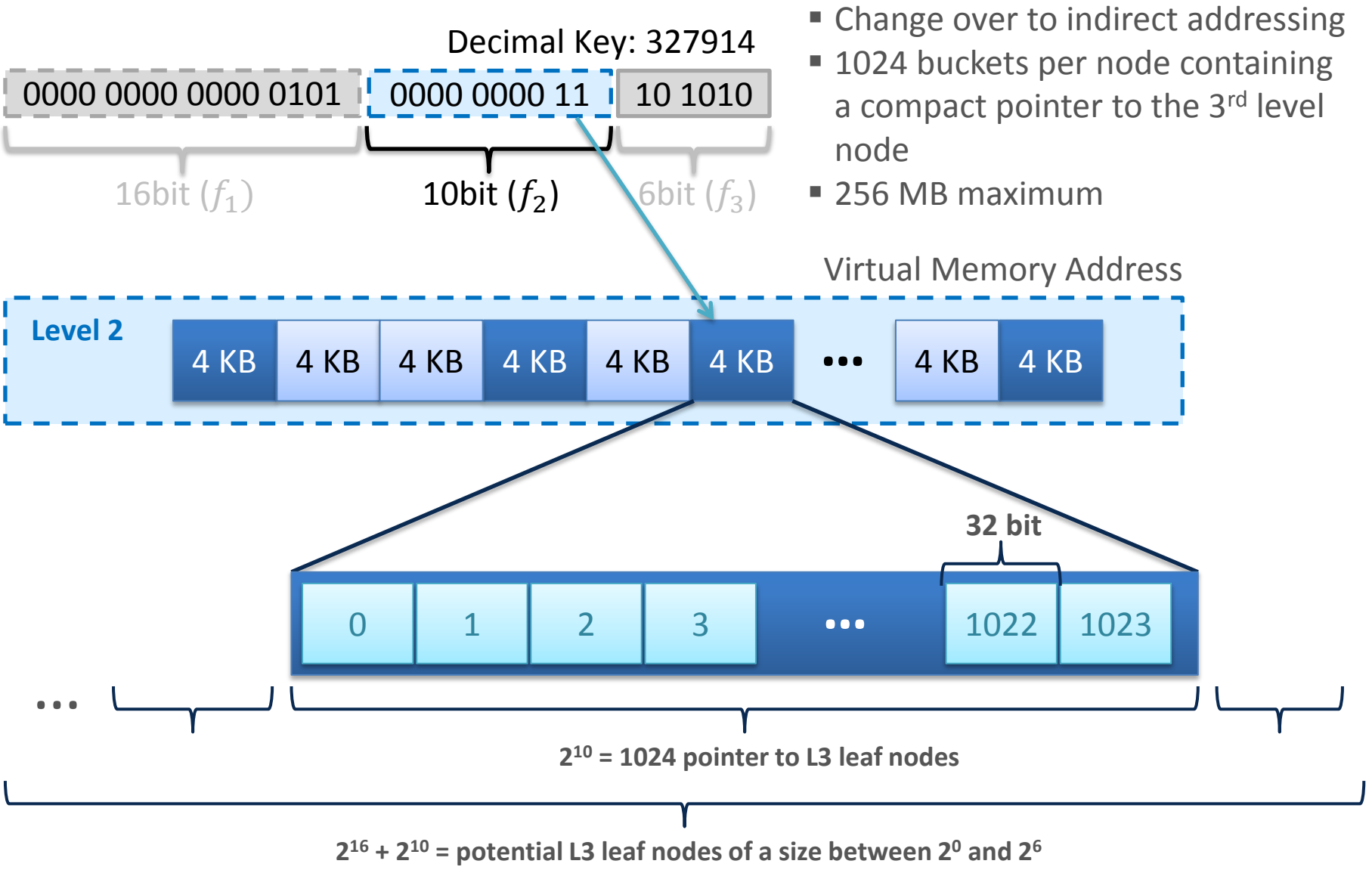
1. INTRODUCTION

indexes) in-memory and use secondary memory (e.g., disks) only for persistence. While disk block accesses constituted the bottleneck for disk-based systems, modern in-memory databases shift the memory hierarchy closer to the CPU and face the "memory wall" [13] as new bottleneck. Thus, they have to care for CPU data caches, TLBs, main memory accesses and access patterns. This essential change also affects indexes and forces us to design new index structures that are now optimized for these new design goals. Moreover, the movement from disks to main memory dramatically increased data access bandwidth and reduced latency. In combination with the increasing number of cores on modern hardware, parallel processing of operations on index structures imposes a new challenges for us, because the overhead



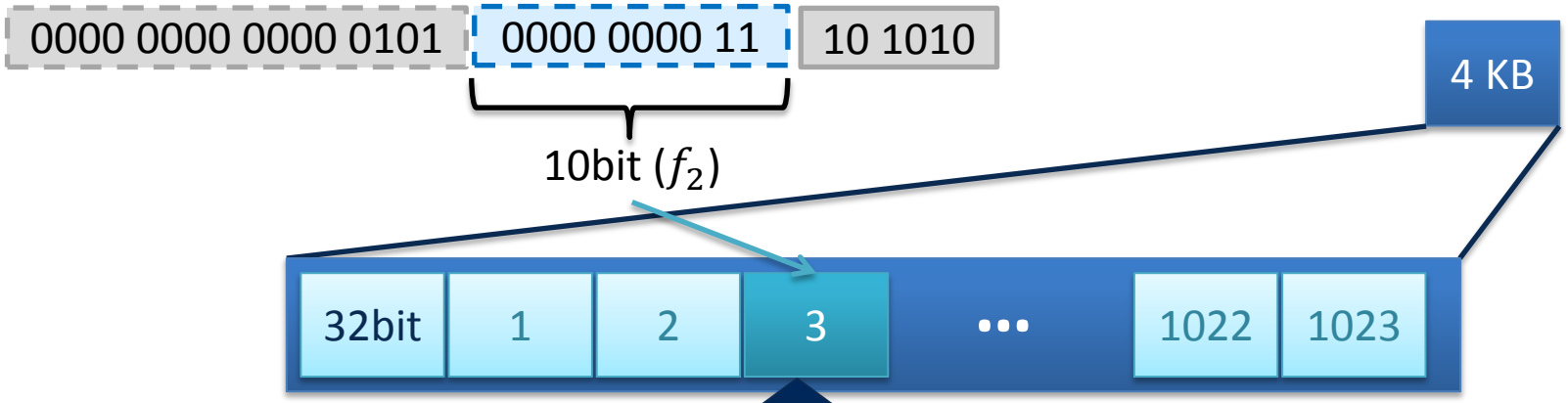
Requirement for Level 2: All nodes have to be stored sequentially in memory

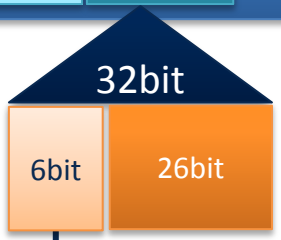

Decimal Key: 327914
0000 0000 0000 0101
0000 0000 11
10 1010
Split key in 3 fragments
16bit (f_1)
10bit (f_2)
6bit (f_3)
Virtual Memory Address
Level 2
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB
Operating System
Page Directory
Memory Management Unit (MMU)
Hardware
Physical Memory Address
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB



- Change over to indirect addressing
- 1024 buckets per node containing a compact pointer to the 3rd level node
- 256 MB maximum

> Level 2: On-demand Level

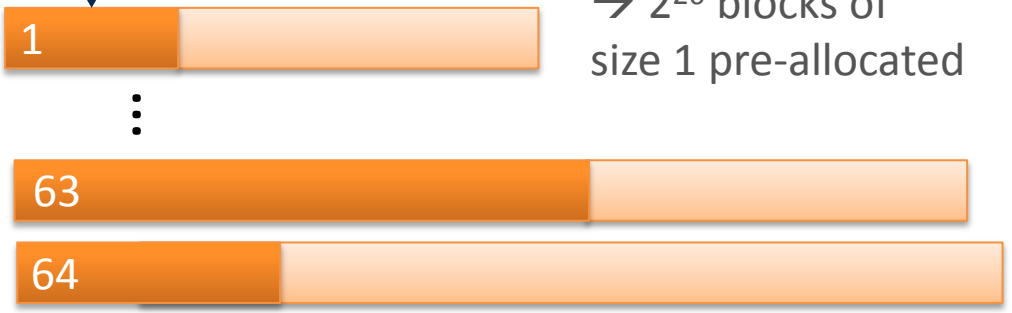




Number of existing elements in L3 leaf node (compression)

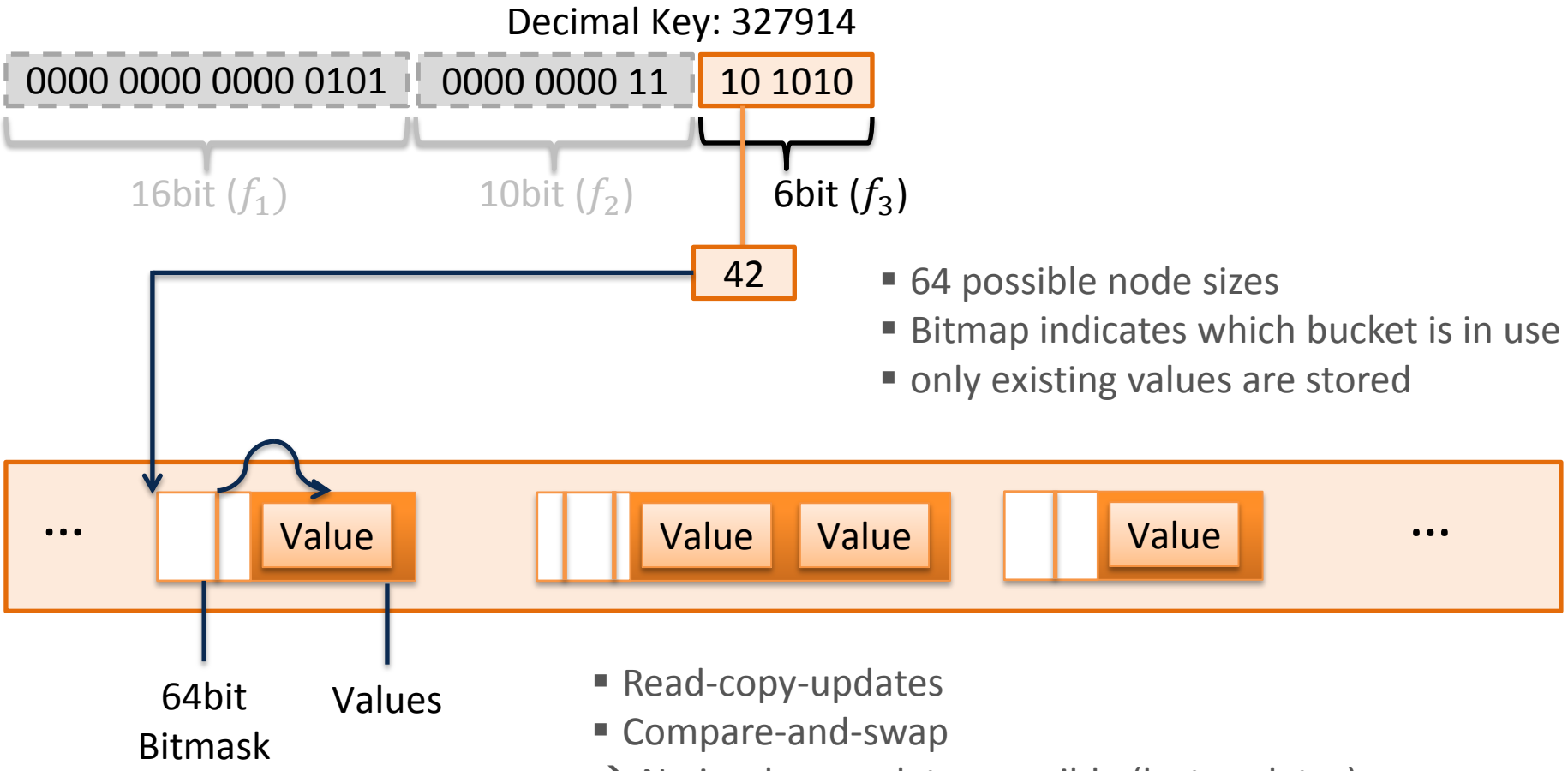
Position of L3 leaf node in corresponding memory segment

- Allocation of consecutive virtual memory segments for each of the 64 node sizes possible on the third level
- Each segment consists of a maximum of 2^{26} blocks of the respective node size



→ 2^{26} blocks of size 1 pre-allocated

→ 2^{26} blocks of node size 64 pre-allocated



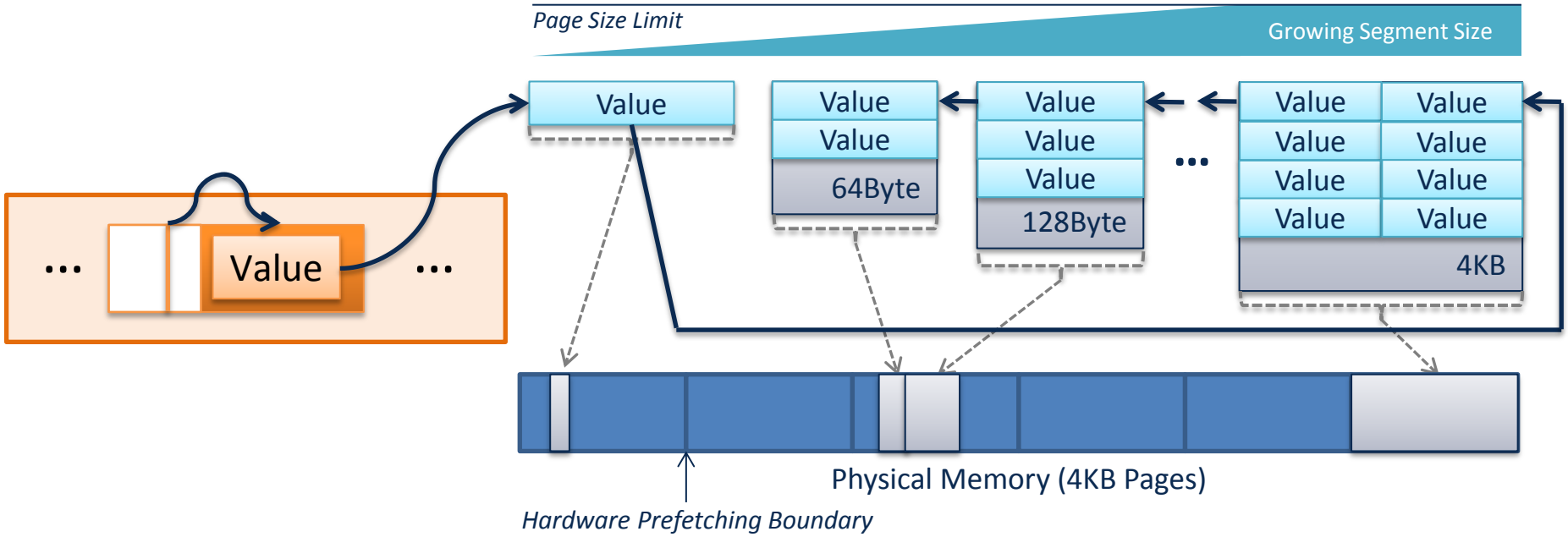
- 64 possible node sizes
- Bitmap indicates which bucket is in use
- only existing values are stored

- Read-copy-updates
- Compare-and-swap

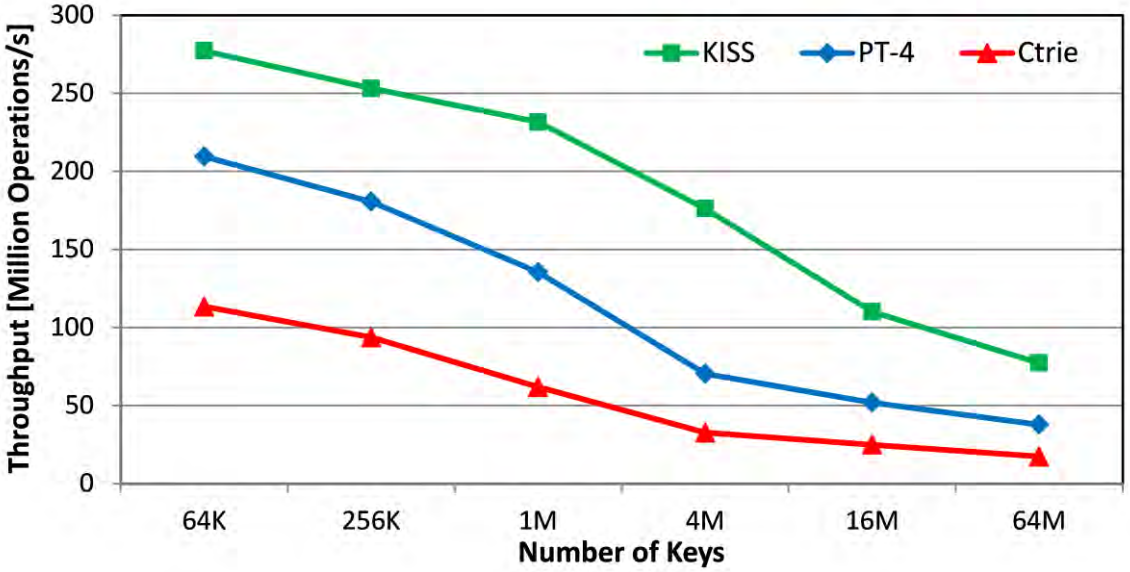
→ No in-place updates possible (lost updates)

Thread-Local Memory Management Subsystem

- Efficient duplicate handling necessary for query processing
  - Scanning a linked list results in random memory accesses



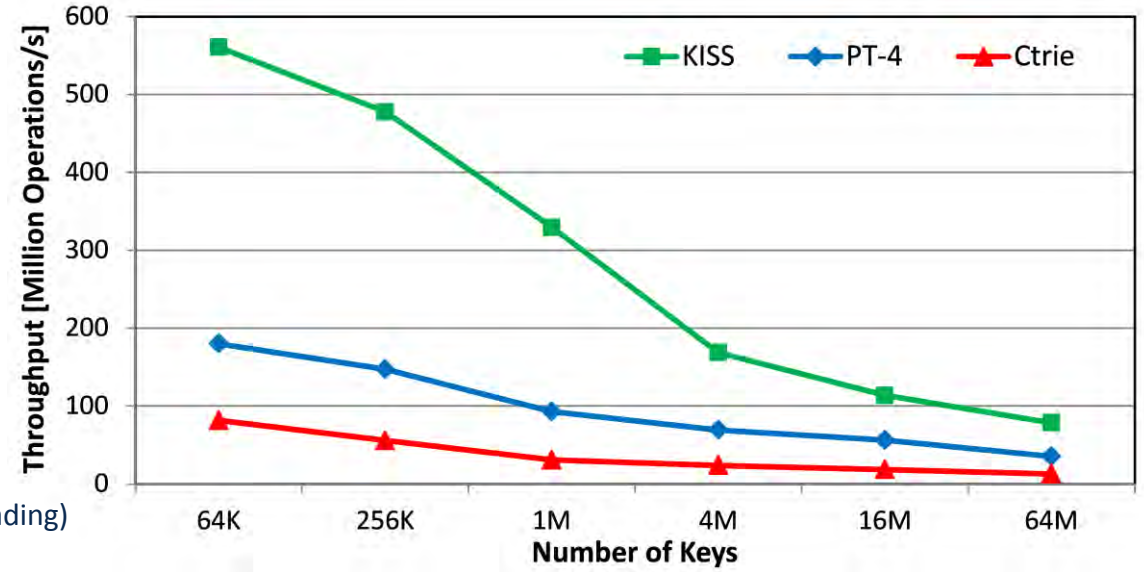
- Page boundaries are a barrier for hardware prefetchers
  - store values sequentially in 4KB blocks
  - blocks grow exponentially until reaching 4KB
    → trade-off between scan performance and memory consumption



Uniform Key Distribution

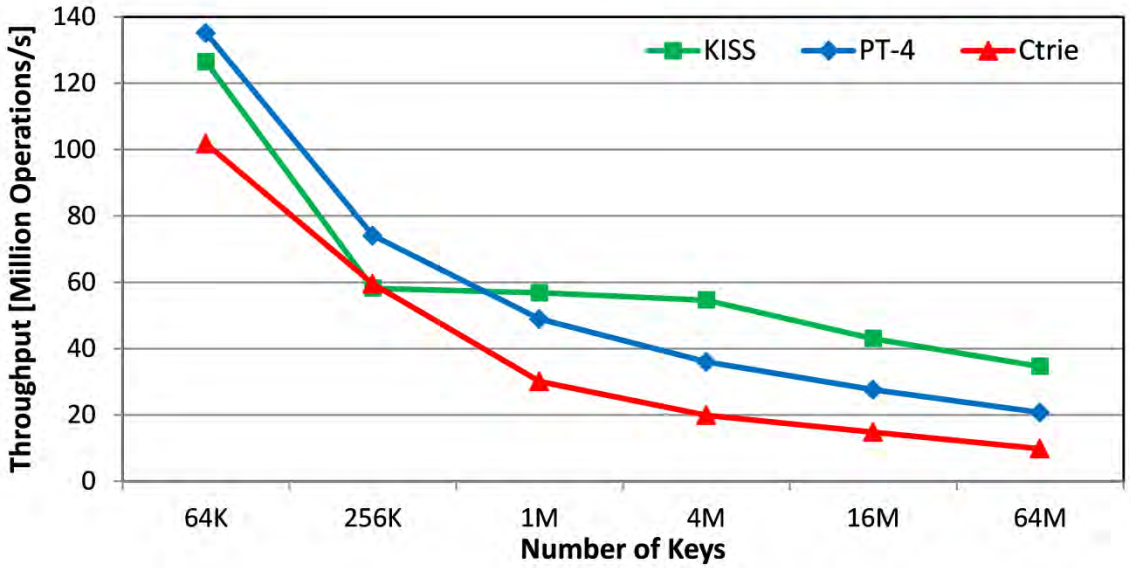
Single-threaded CSB⁺-Tree: ~3Million Ops/s
→ KISS-Tree: 18 Million Ops/s with one thread
(64M Keys)

Sequential workload



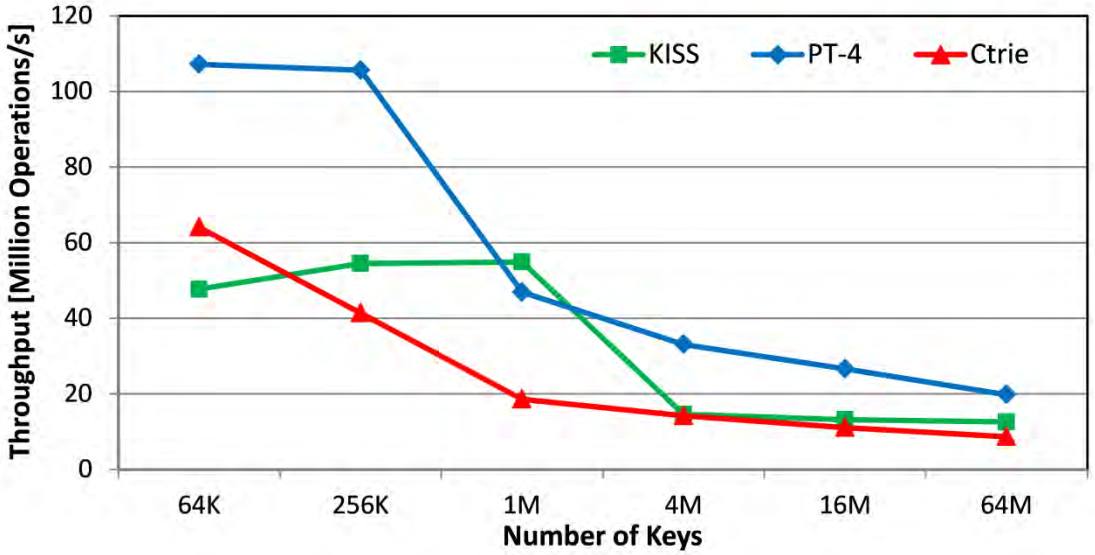
Workloads
- 8 Byte Payload

Evaluation Hardware
- Intel i7-2600 (SMP, 4 cores with Hyper-Threading)
- 8 MB LLC, 16 GB main memory



Uniform Key Distribution

Sequential workload



Workloads
- 8 Byte Payload

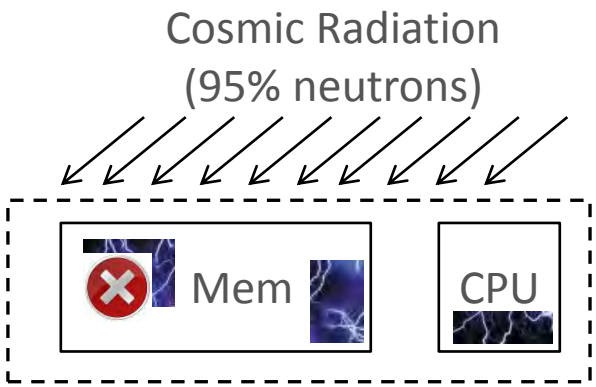
Evaluation Hardware
- Intel i7-2600 (SMP, 4 cores with Hyper-Threading)
- 8 MB LLC, 16 GB main memory

Some Techniques

- Compression
- Indexing
- Resilience
- (Hardware) Transactional Memory

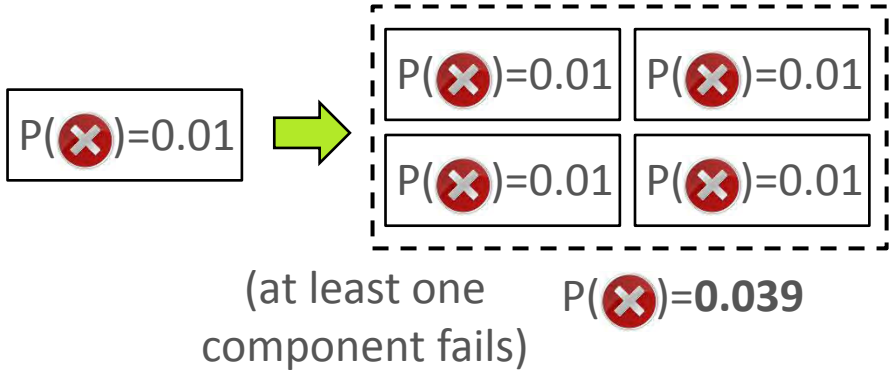
Increasing Component Error Rates

- Decreasing feature sizes (new tech generations)
- Reduced voltage supply
- Static (hard) vs. **dynamic (soft) errors**
- **8% increase** error rate per tech generation **[Borkar05]**
- **25,000 – 70,000** FIT / Mbit **[Schroeder09]**



Increasing System Error Rates

- Increasing scale
  - # of components (core, transistor)
  - Memory capacities
- Example:
  - Fixed error rate / component



➔ ***Errors and error-prone behavior will become the normal case***

Implicit (silent) vs. Explicit (detected/corrected) Errors

- State-of-the-art: error detection and correction at HW/OS level

*State-of-the-Art: **Resilient Memory***

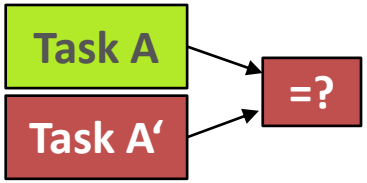
- ECC / parity bits / memory scrubbing / full data redundancy

ECC Extended Hamming(7+1,4)

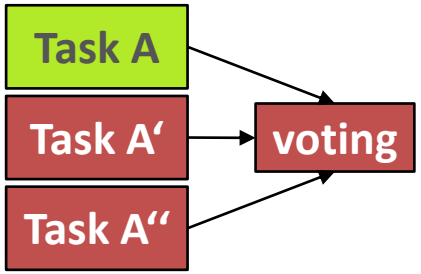
*State-of-the-Art: **Resilient Computing***

- Computation redundancy

Double Modular Redundancy (DMR):



Triple Modular Redundancy (TMR):



➔ ***Such resiliency mechanisms cause „resiliency costs“***

Resiliency Costs Categories

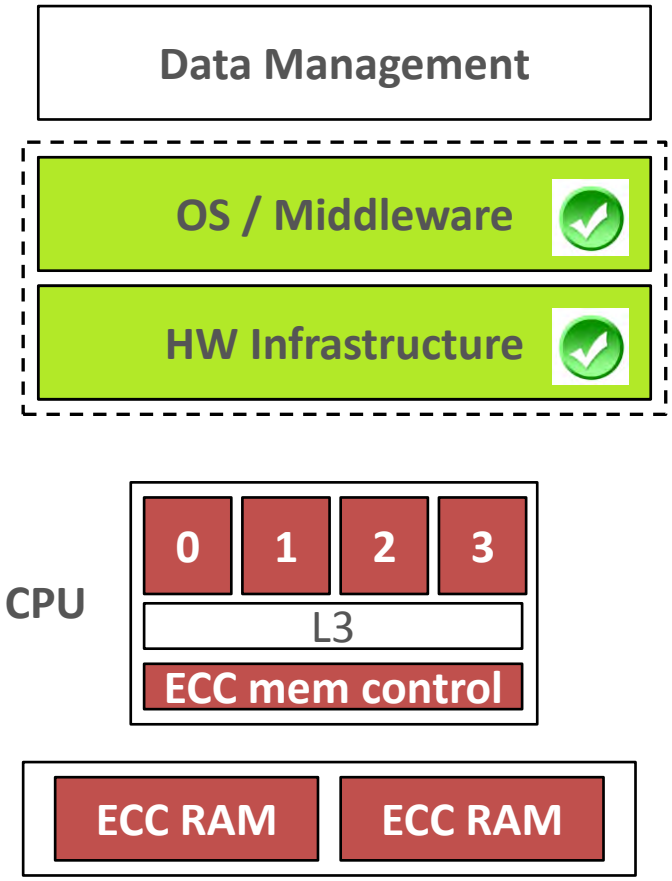
- Performance overhead (throughput, latency)
- Memory overhead
- Energy consumption
- Monetary HW costs

*Resiliency Costs @ **OS-Level***

- **Memory overhead** (capacity, bandwidth)
- **Computation overhead**
- Energy consumption (increased time)

*Resiliency Costs @ **HW-Level***

- **Monetary HW costs** (Chipset, ECC RAM)
- **Energy consumption** (time, chip space)
- Computation overhead



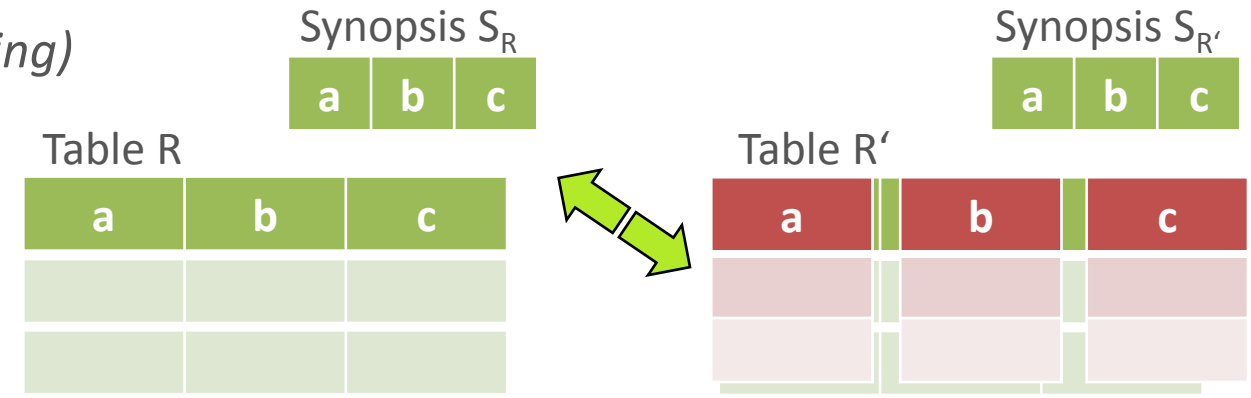
➔ ***Increasing error rates ~ increasing resiliency costs!***

Challenge

- Problem: data loss/corruption (explicit/implicit)
- Goal: data stability by data redundancy and error correction

Example (Data Partitioning)

- Table R (a,b,c)
- Data redundancy (synopsis and replicas)



Test Scheduling
Multiple Replicas
Workload Characteristics

Time-based /on-the-fly error detection and correction

Optimization

- Exploit the multiple replicas ➔ **(complementary)** layouts
- E.g., different sorting orders, partitioning schemes, compression schemes, etc

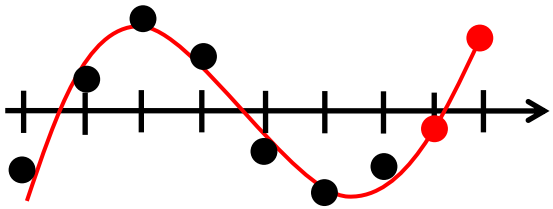
Challenge

- Problem: missing/invalid tuples (explicit/implicit)
- Goal: reliable query results by error correction / error-tolerant algorithms

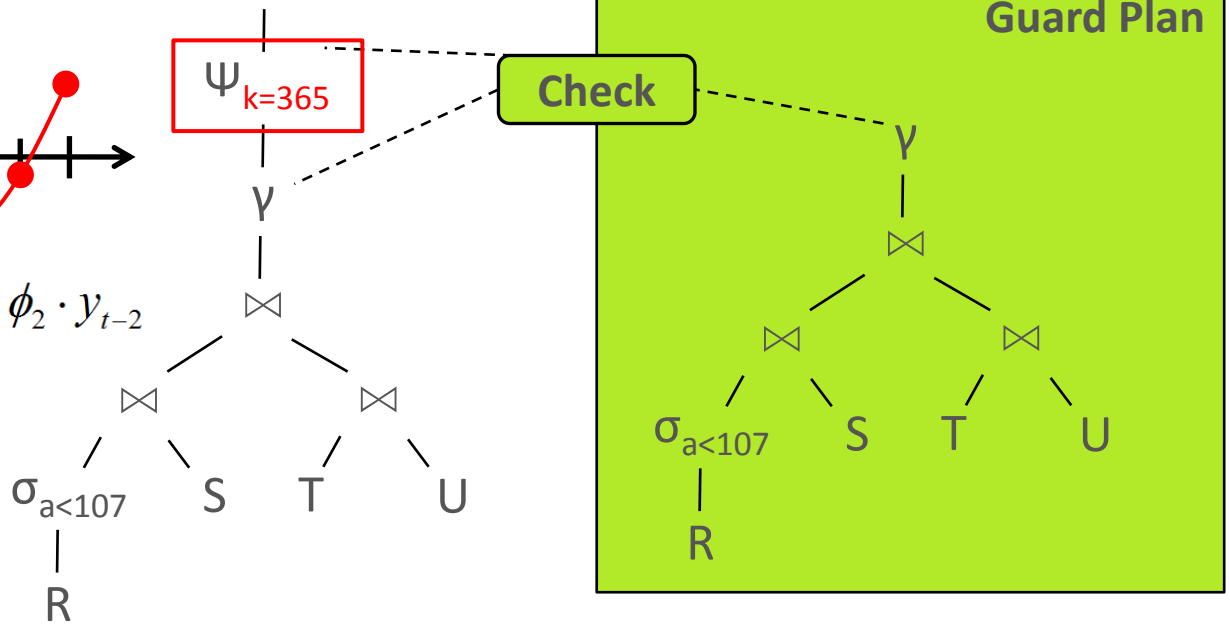
Plan Scheduling
Operator Semantics
Intermediate Results

Example (Advanced Analytics)

- Q: $\Psi_{k=365}(\gamma(\ \sigma_{a<107}R \bowtie S \bowtie T \bowtie U\ ))$
- Computation redundancy

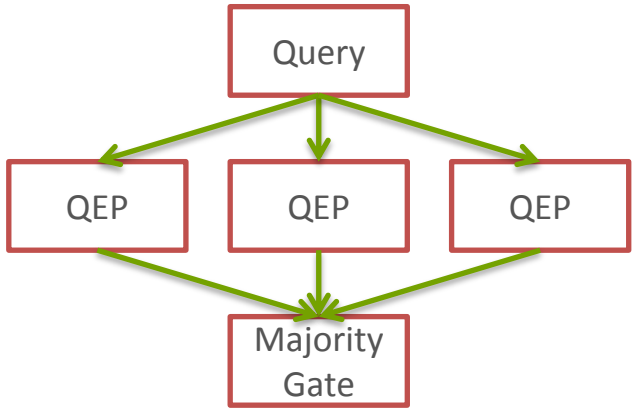


$$AR(2): \hat{y}_t = \phi_1 \cdot y_{t-1} + \phi_2 \cdot y_{t-2}$$



Redundant Query Processing (Triple Modular Redundancy)

- Based on a single query (→ three times execution of an optimal QEP)
- Majority Gate on a single tuple basis
- Drawbacks
  - Assumption of valid raw data (single point of truth)
  - Corrupt raw data are propagated through the model

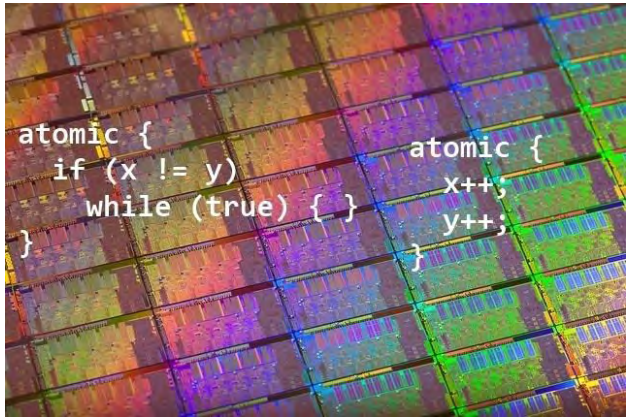


Alternative

- Based on a single query (→ three different QEPs)
- Majority Gate on a single tuple basis
- Different QEPs based on
  - different operators and
  - in particular on redundant data sources
- Properties
  - no single point of truth
  - data replication is important

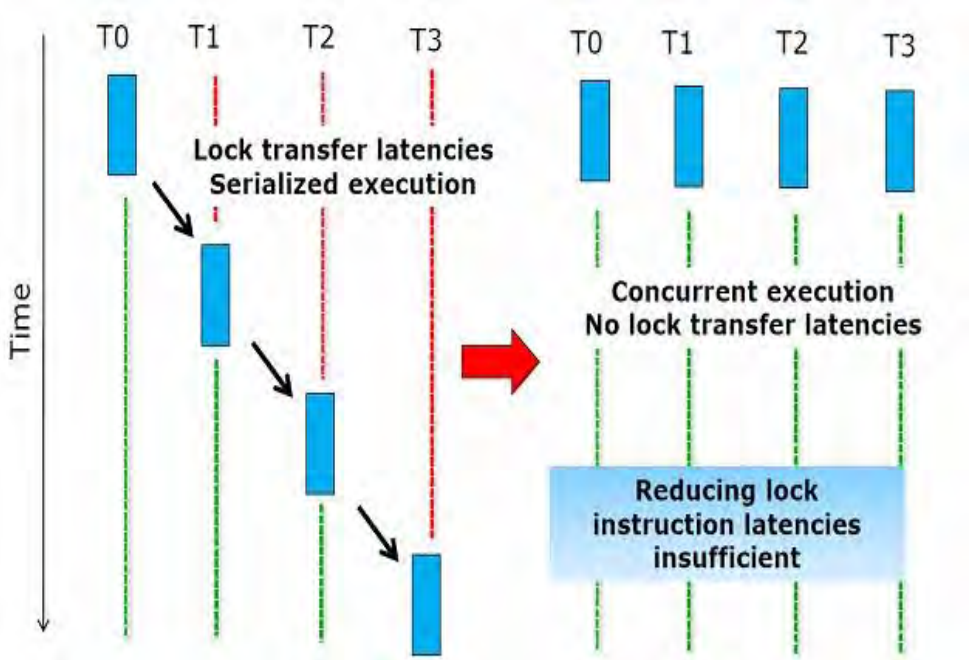
Some Techniques

- Compression
- Indexing
- Resilience
- (Hardware) Transactional Memory

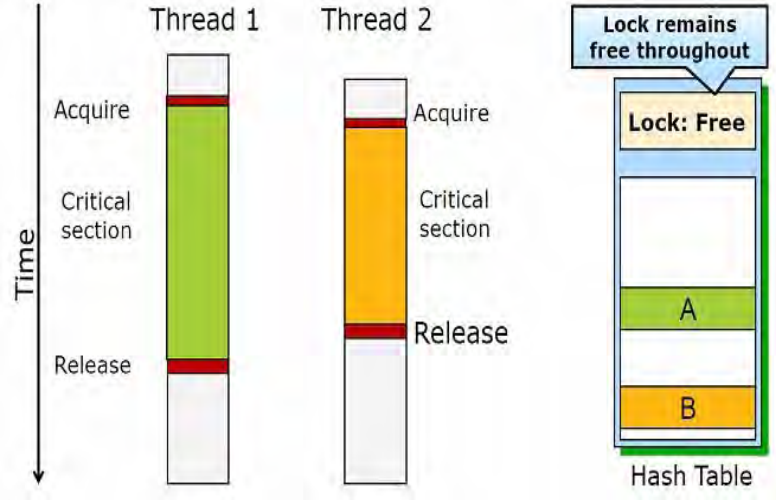


Concurrent Execution

- No serialization, no communication if no data conflicts
- Data conflicts are atomatically termined by the underlying TM infrastructure



Example: Hash Table

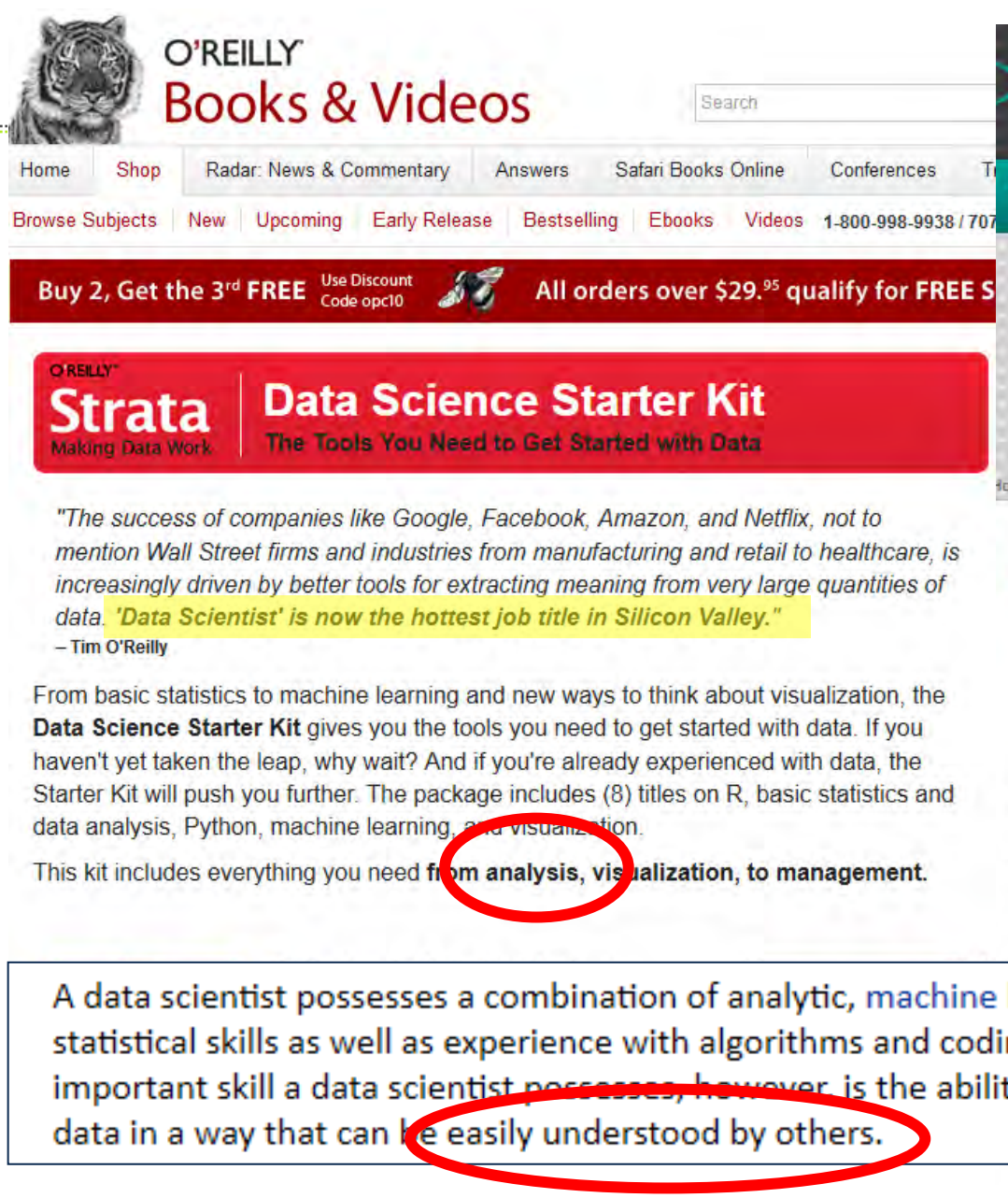


Main Memory Centric Data Management

What are the main challenges?

Technology perspective

Application perspective



A data scientist possesses a combination of analytic, machine learning, data mining and statistical skills as well as experience with algorithms and coding. Perhaps the most important skill a data scientist possesses, however, is the ability to explain the significance of data in a way that can be easily understood by others.

let us map the situation of data analytics to ...

Phillip G. Armour

The Five Orders of Ignorance

Viewing software development as knowledge acquisition and ignorance reduction.

0th Order Ignorance (0OI)—*Lack of Ignorance*

- I know something and can demonstrate my lack of ignorance (0OI is knowledge)
- When I have 0OI, I have the answer to the problem.

Mapping to data analytics

- Data Consumption Side
  - no need to look at the data
- Data Cube / Data Space Selection
  - no need to identify, tap, and combine potentially interesting data cubes
- Data Provisioning
  - don't bother about data analytics

1st Order Ignorance (1OI) *— Lack of Knowledge*

- I don't know something and can readily identify that fact (1OI is basic ignorance)
- When I have 1OI, I have the question.
- Usually, having a good question makes it fairly easy to find the answer

Mapping to data analytics

- Data Consumption Side
  - Reporting!
    you know the dimension values
    you are looking specifically for a measure of the specific dimension values
- Data Cube / Data Space Selection
  - no need to actively search for possibly interesting data cubes
    place where to search your information is well known
- Data Provisioning
  - data sources are well-known and understood; integration/transformation steps exist

2nd Order Ignorance (2OI) *— Lack of Awareness*

- I don't know that I don't know something
  (not only am I ignorant of something, I am unaware of this fact. I don't know enough to know that I don't know enough.)
- Not only do I not have the answer I need, I don't even have the question.

Mapping to data analytics

- Data Consumption Side
  - Guided Navigation / Explorer-Style
- Data Cube Selection
  - ??? how can I find the ‚right' data
- Data Provisioning
  - I don't know my sources, but I suspect that there is helpful data out there...

3rd Order Ignorance (3OI)— Lack of Process

- I have 3OI when I don't know a suitably efficient way to find out I don't know that I don't know something → lack of process
- If I have 3OI, I don't know of a way to find out there are things I don't know that I don't know

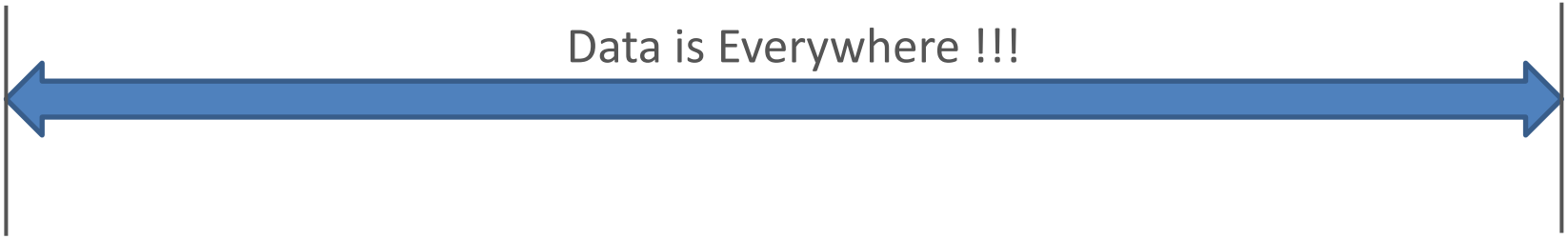
Mapping to data analytics

- Data Consumption Side
  - Data mining („Tell me something interesting!")
  - Data visualization
- Data Cube Selection
  - I don't even know how to design my BI entities
- Data Provisioning
  - I don't know anything about my data sources

Main Memory Centric Data Management

What are the main benefits?

Help business people with an ***xth Order Ignorance***

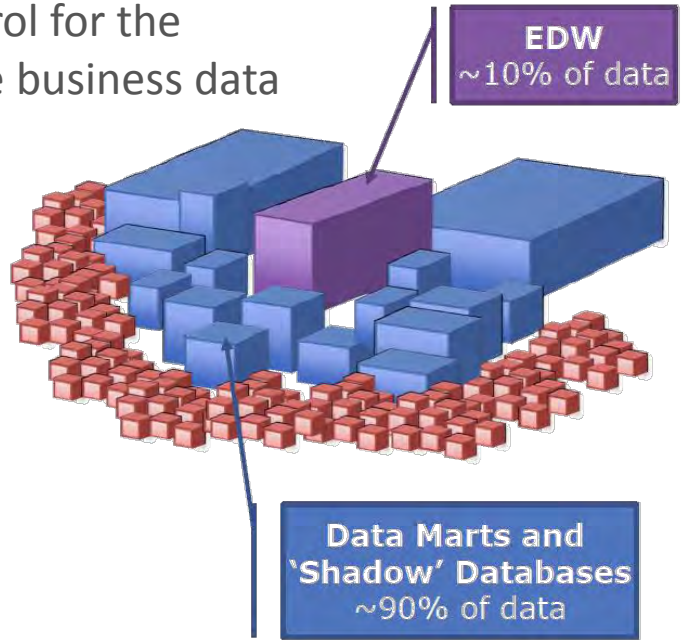


Corporate EDW(s)

- IT governed infrastructure
- well defined with standardized processes, taxonomies, and KPIs
- Top-down control for the enterprise-wide business data

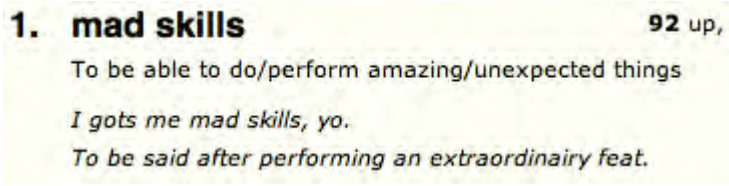
Dozens of data marts, 100s of local databases, 1000s of spreadsheets

- locally defined analytics
- Bottom-up generation and ad-hoc access pattern



Extend DWH Infrastructure

Self-Service/ Agile BI



Magnetic

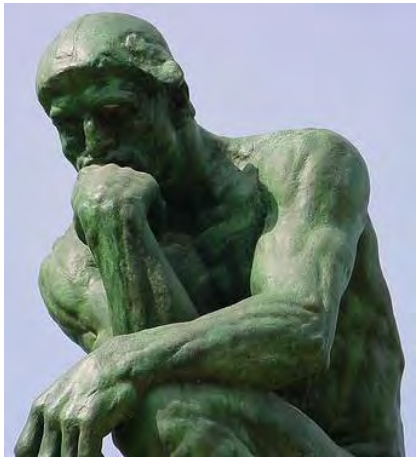
- **„Attract data and practitioners“**
- Usage of all data source independet of their data quality

Agile

- **„Rapid iteration: ingest, analyze, productionalize“**
- Continous evolution of the logical and physical structures
- ELT (Extraction, Loading, Transformation)

Deep

- **„Sophisticated analytics in Big Data“**
- Extended algorithmic run-time
- Ad-hoc advanced analytics and statistics



Requirement 1: Balance Performance and Data Volume

- High query performance for interactive analysis / data exploration
- **Challenge:** Huge number of parallel users and ad-hoc query/data schemes

- Batch processing for offline tasks
  (hypothesis generation / consistency checks following complex business semantics)
- **Challenge:** Need to massage massive data volumes with complex business logic

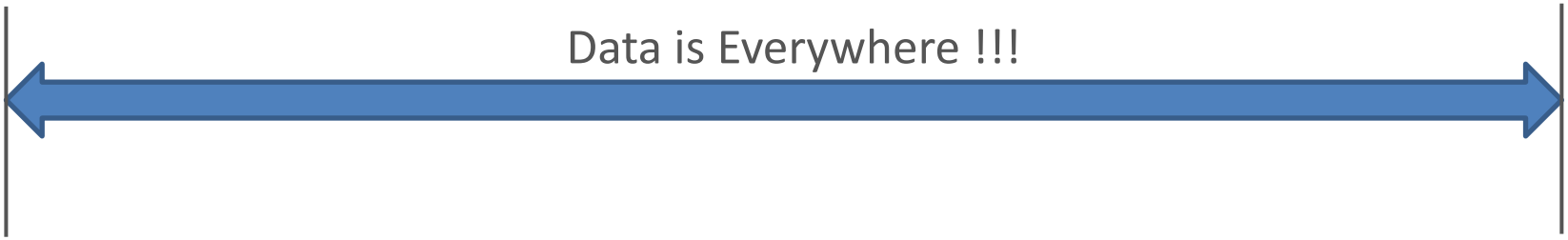
Requirement 2: Support Pull and Push Processing

- Full spectrum of load granularity (batch – trickle feed – stream processing)
- **Challenge:** Need to integrate data streaming systems into DWH infrastructure

Requirement 3: Corporate Memory

- „Expect the unexpected“
- **Challenge:** Manage the organization’s archives and individuals’ memories

- Exist in hybrid environments
- **Challenge:** Provide cloud-enabled DWH environments

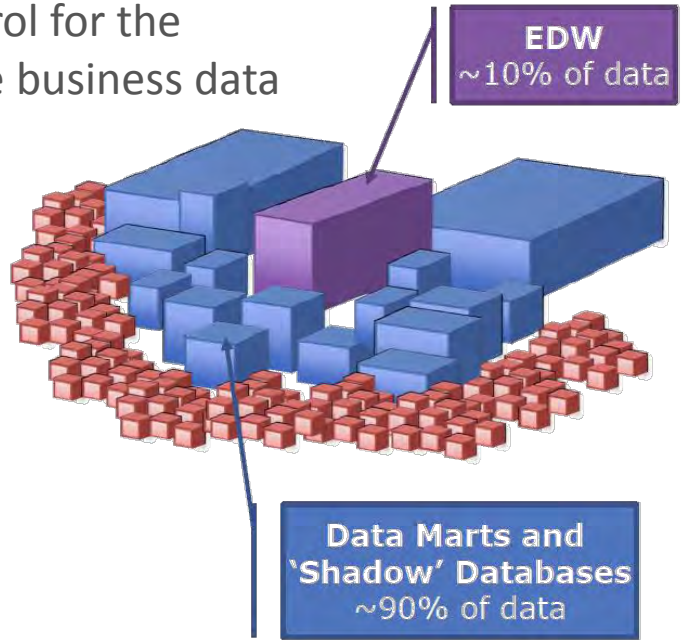


Corporate EDW(s)

- IT governed infrastructure
- well defined with standardized processes and measures
- Top-down control for the enterprise-wide business data

Dozens of data marts, 100s of local databases, 1000s of spreadsheets

- Locally defined analytics
- Bottom-up generation and ad-hoc access pattern



Extend DWH Infrastructure

Self-Service/ Agile BI

Approach: Enable business analysts to

- For business expert users, provide infrastructure to create agile data spaces
- For experienced users, easily run **ad-hoc queries** on complex data warehouses

Challenges

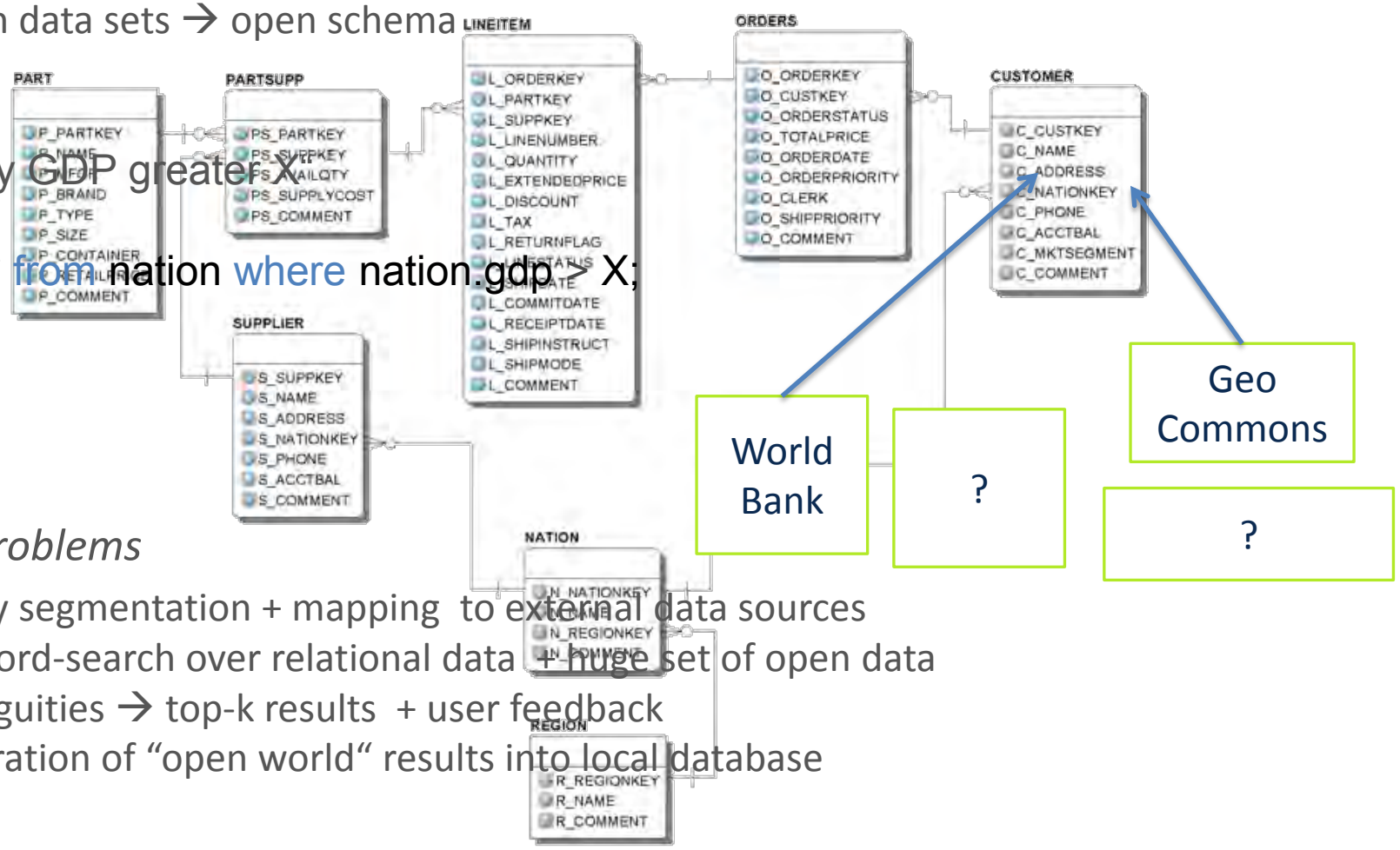
- Find the right **input language** for business people (**simplified natural language**)
- Work on **very large and complex** database schemas (hundreds of tables, inheritance patterns)
- **Incrementally** improve meta-model by experience gained from previous ad-hoc queries
- Balance well-structured and IT-governed data flow with ad-hoc-analytics requirements
- Reduce time to consumption – no time left for tuning mechanisms

„Drill Beyond“: Extending OLAP using Open Data

- Open World“ approach: Include all external open data sets → open schema

„country GDP greater X“

```
select * from nation where nation.gdp > X;
```

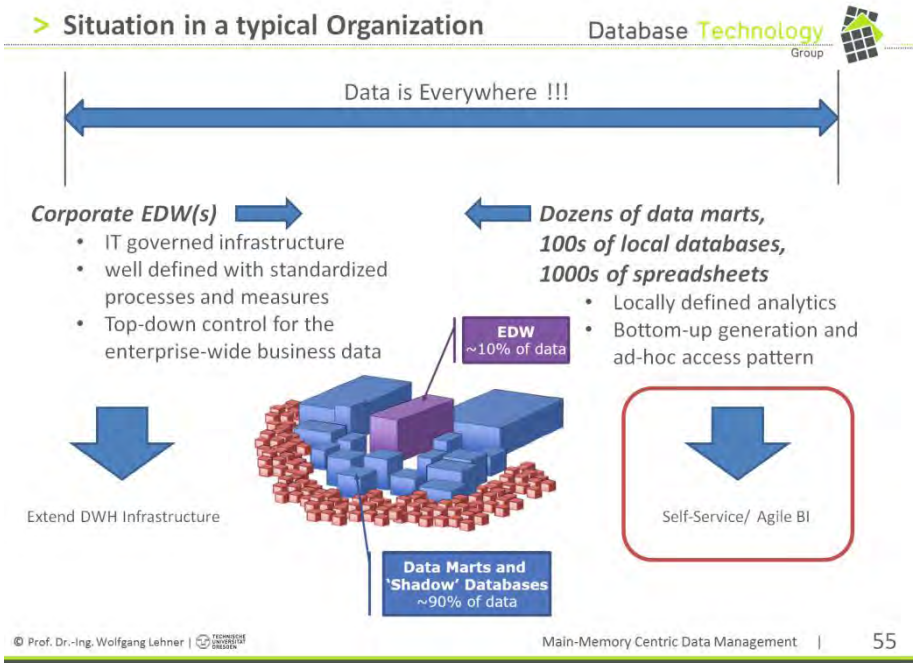


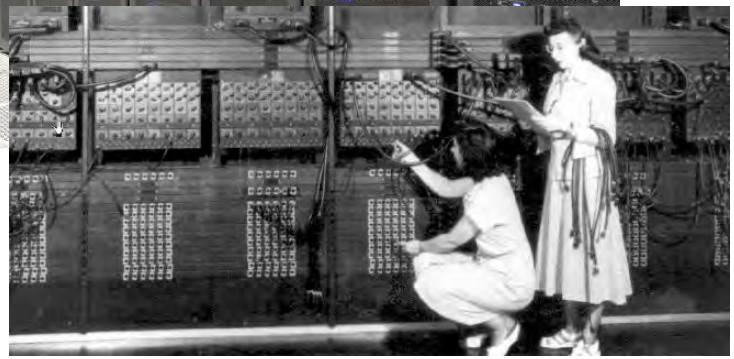
Main Problems

- Query segmentation + mapping to external data sources
- Keyword-search over relational data + huge set of open data
- Ambiguities → top-k results + user feedback
- Integration of “open world“ results into local database

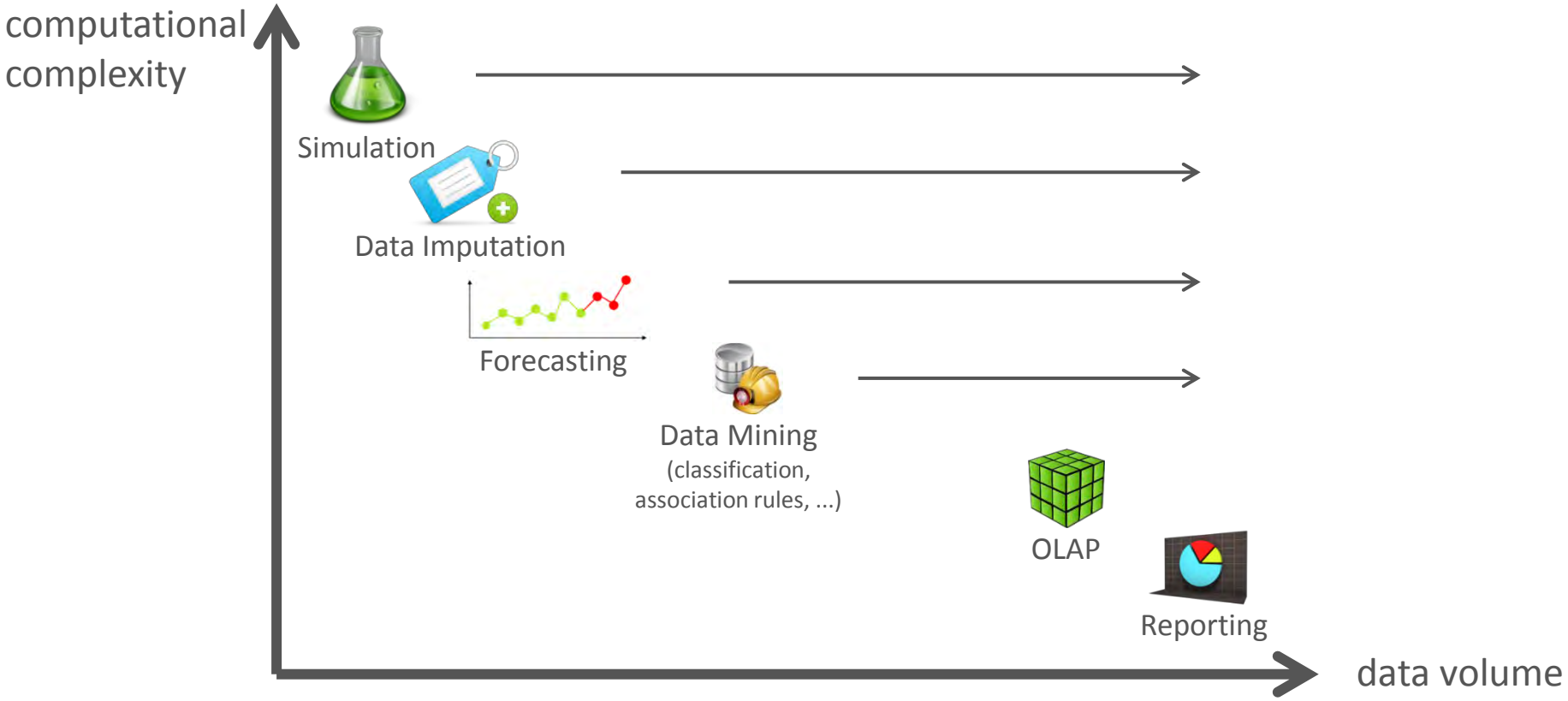
Data Management

High Performance Computing





data crunching *meets* ***number crunching***



NETFLIX

Member Sign In

Start Your 1 Month Free Trial | How It Works | Browse Selection | 1 Month Free Trial Info

Watch as many movies as you want! For only $7.99 a month.

- ✓ Streaming instantly over the Internet to your PC, Mac & TV
- ✓ Only $2 more a month to get unlimited DVDs by mail
- ✓ Cancel anytime

Questions? 1-866-636-3076
24 hours a day

We're not sure you will be able to sign up for Netflix from your area.

You will need a valid U.S. mailing address to sign up for Netflix. Also, you will only be able to watch instantly if you are in the 50 United States or Washington, D.C. It looks like you are outside the United States. If this is incorrect, please contact your Internet provider for help. We are sorry for any inconvenience.

I have a valid U.S. mailing address

Unlimited TV episodes & movies insta

Connect devices like these to your Netflix account to **watch instantly on your TV.**

Wii | PS3 | XBOX 360

Learn more › | Learn more › | Learn more ›

Watch as often as you want, anytime you want.

PLUS

See other devices that stream instantly from Netflix

> The NetFlix Competition (2)

Netflix' star rating system helps determine personalized movie recommendations. Now the company is looking to outside developers to improve those recommendations.

BUSINESS

The $1 Million Netflix Challenge

FRIDAY, OCTOBER 6, 2006 | BY KATE GREENE

VP Jim Bennett discusses how recommendation systems suggest your next movie and the challenges of building a better one.

E-mail Audio » Print

Earlier this week, Netflix, the online movie rental service, announced it will award $1 million to anyone who can come up with an algorithm that improves the accuracy of its movie recommendation service.

In doing so, the company is putting out a call to researchers who specialize in mac learning--the type of artificial intelligence used to build systems that recommend m books, and movies. The entrant who can increase the accuracy of the Netflix recommendation system, which is called Cinematch, by 10 percent by 2011 will w prize.

Recommendation systems such as those used by Netflix, Amazon, and other Wel retailers are based on the principle that if two people enjoy the same product, they' likely to have other favorites in common too.

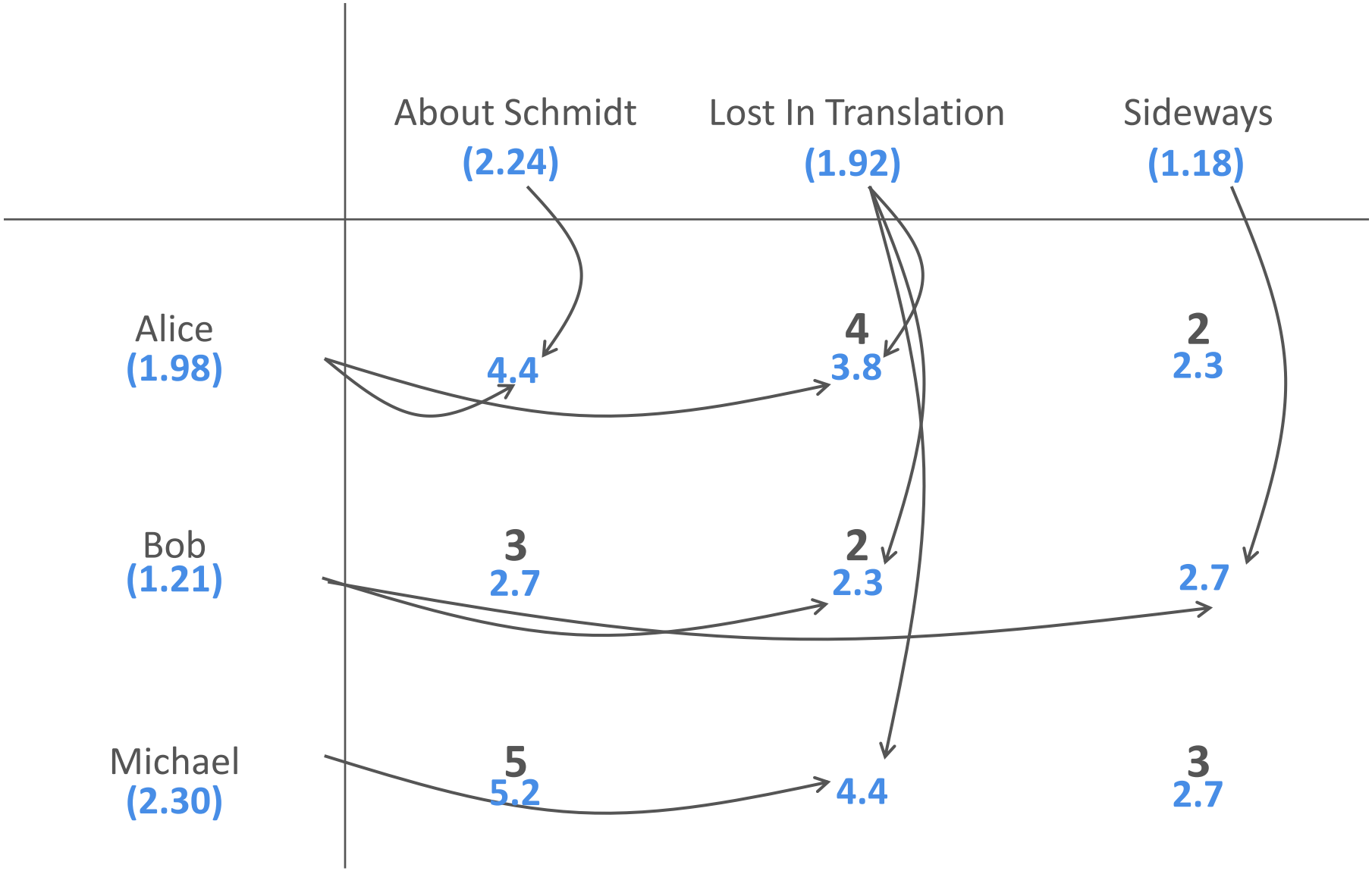
But behind this simple premise is a complex algorithm that incorporates millions of ratings, tens of thousands of items, and ever-changing relationships between user preferences.

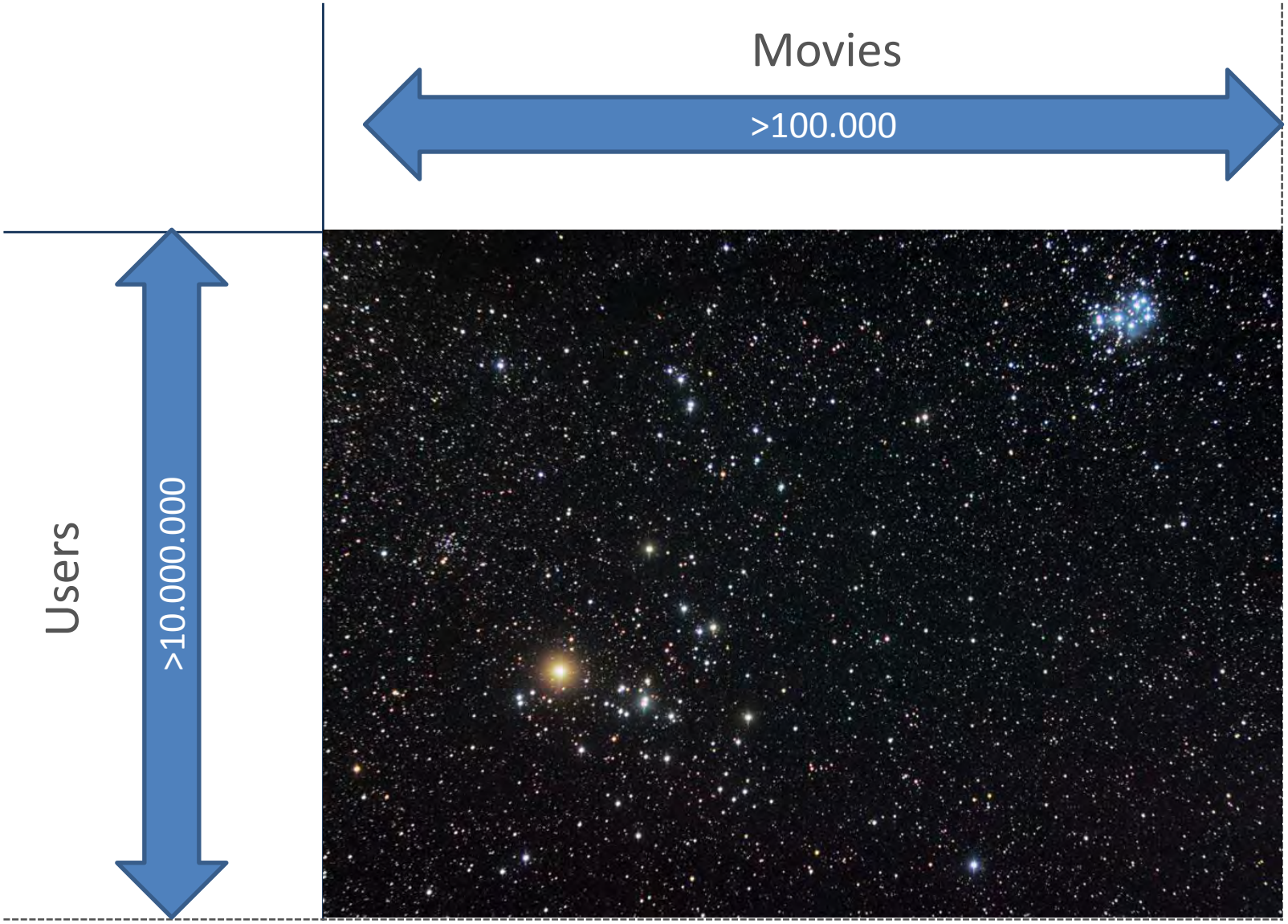
BellKor's Pragmatic Chaos

$$\hat{r}_{ui} = b_{ui} + |\mathrm{N}(u)|^{-\frac{1}{2}} \sum_{j \in \mathrm{N}(u)} e^{-\beta_u \cdot |t_{ui} - t_{uj}|} c_{ij} +$$

$$|\mathrm{R}(u)|^{-\frac{1}{2}} \sum_{j \in \mathrm{R}(u)} e^{-\beta_u \cdot |t_{ui} - t_{uj}|} ((r_{uj} - \tilde{b}_{uj}) w_{ij}) +$$

$$\sum_{j \in \mathrm{R}(u)} e^{-\gamma_u \cdot |t_{ui} - t_{uj}|} ((r_{uj} - \tilde{b}_{uj}) d_{ij}) .$$


About Schmidt
(2.24)
Lost In Translation
(1.92)
Sideways
(1.18)
Alice
(1.98)
4.4
4
3.8
2
2.3
Bob
(1.21)
3
2.7
2
2.3
2.7
Michael
(2.30)
5
5.2
4.4
3
2.7


Movies
>100.000
Users
>10.000.000

color code := user rating

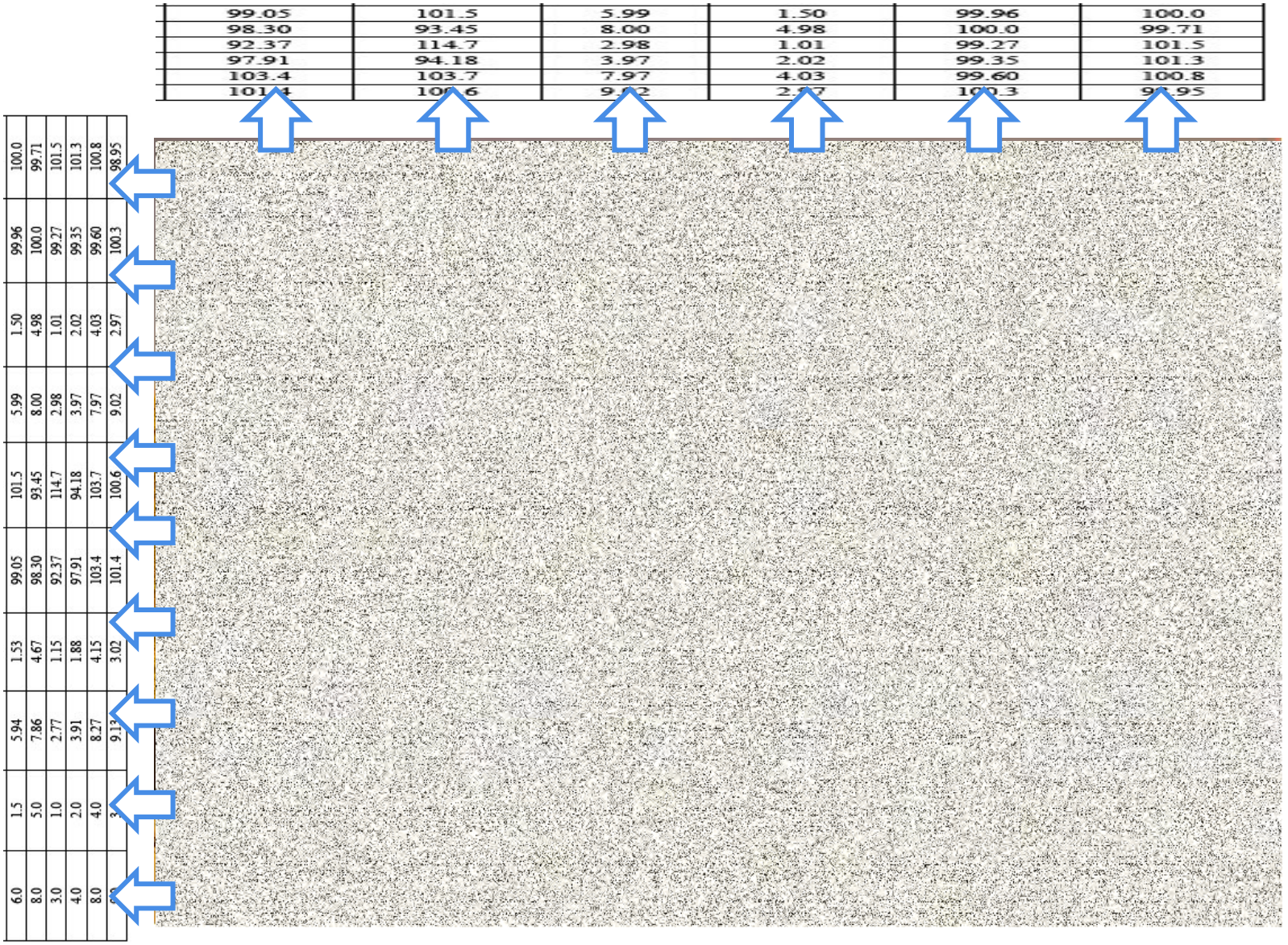
different movies

different users

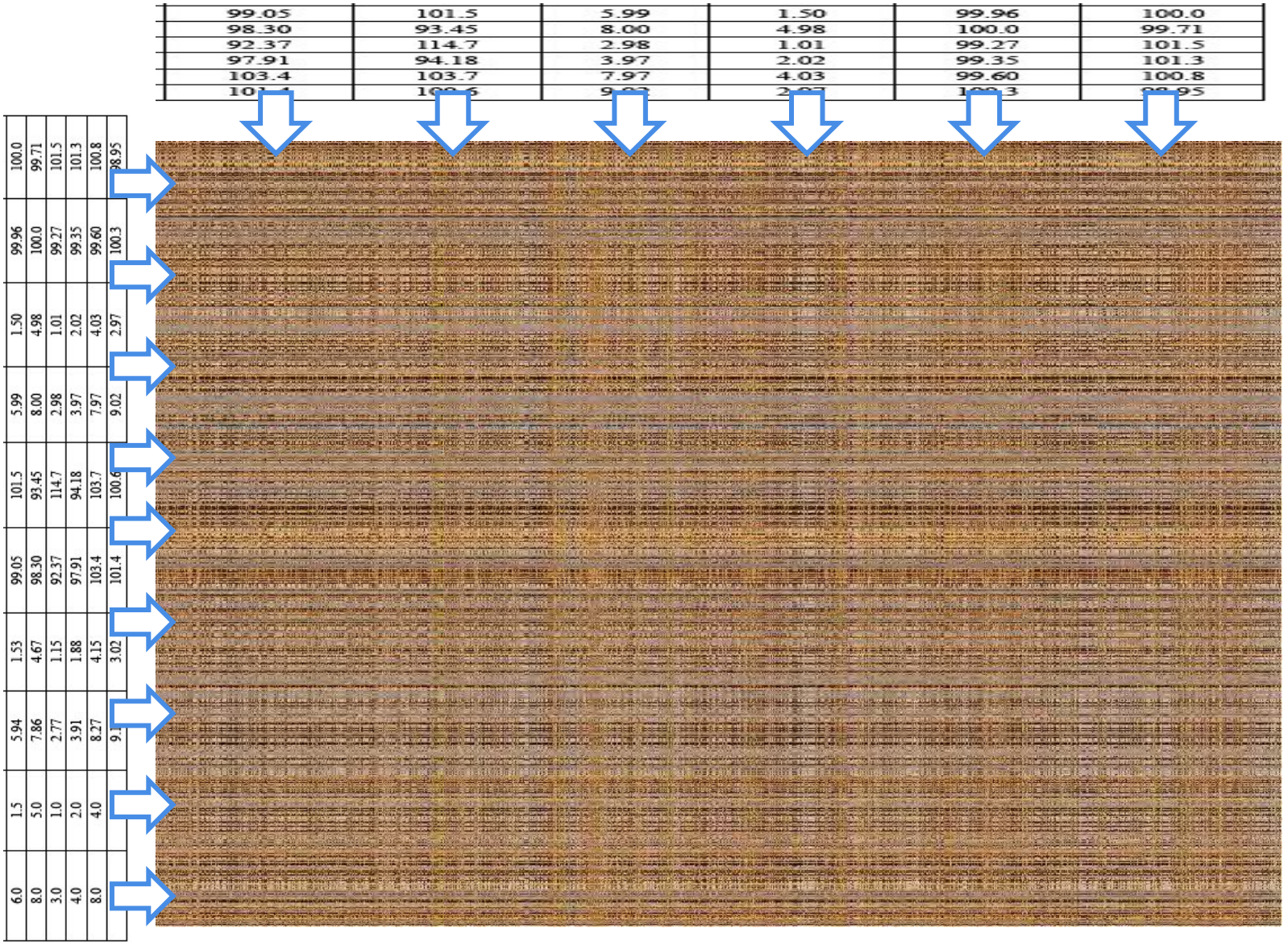
Phase 1: drop 75% of all pixels

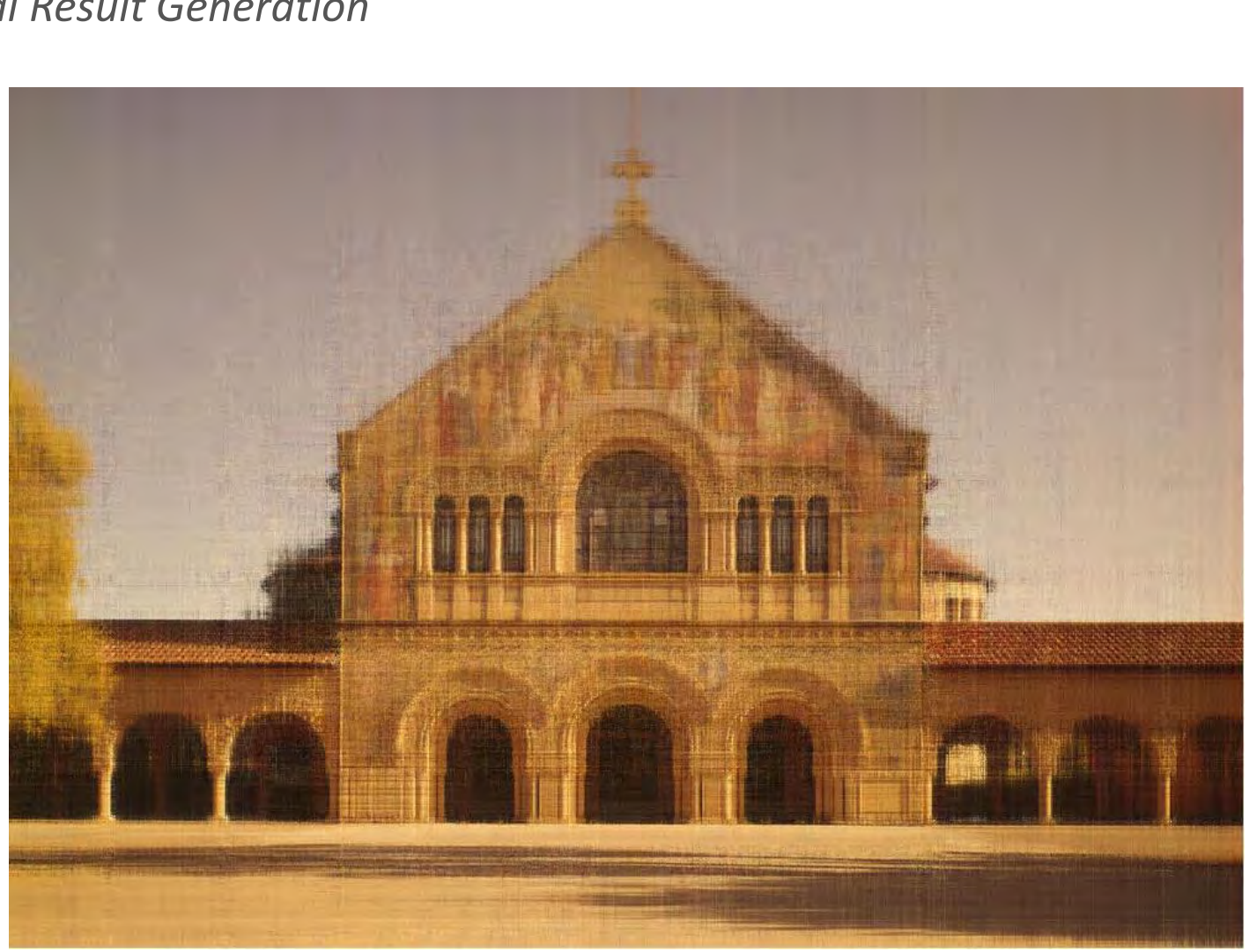
Phase 2: Random permutation of rows and columns

Phase 3: Determine the latent factors

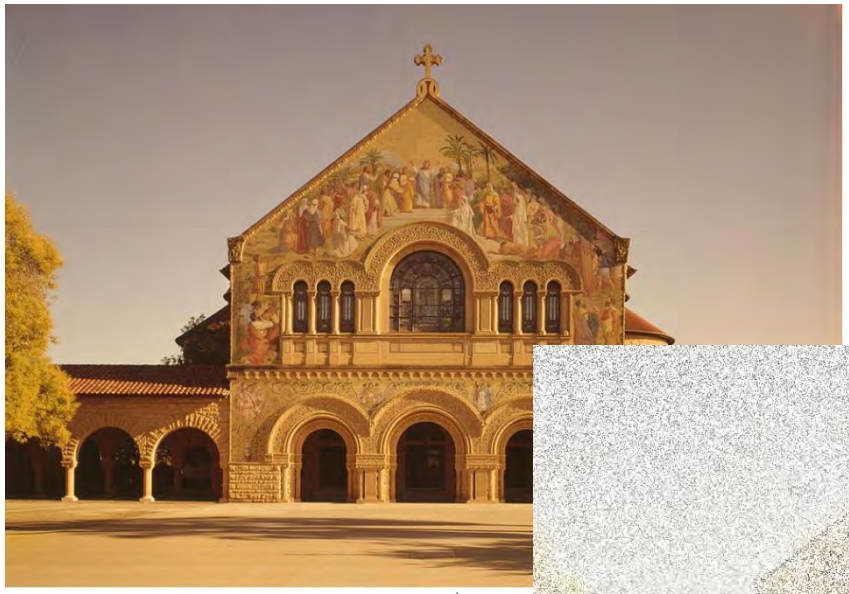


Phase 4: Reconstruction





Phase 5: Final Result Generation



25%

LFM

The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data.

Welcome to the Petabyte Age.

Main Memory Centric Data Management

- requires a holistic picture from system level to end-user experience
- data management layer is still the name of the game
- novel applications will combine the complexity of number-crunching and data-crunching

Huge impact on

- design of analytics infrastructures
- design of database systems
- implementation of analytical applications

to finally conclude: Recap

- 5 Orders of Ignorance
  - 0OI - Lack of Ignorance
  - 1OI - Lack of Knowledge
  - 2OI - Lack of Awareness
  - 3OI - Lack of Process
  - 4OI?

4th Order Ignorance (4OI)— Meta Ignorance

I have 4OI when I don't know about the Five Orders of Ignorance.

... we hopefully passed that stage!

Main-Memory Centric Data Management – Open Problems and Some Solutions

Wolfgang Lehner

Technische Universität Dresden