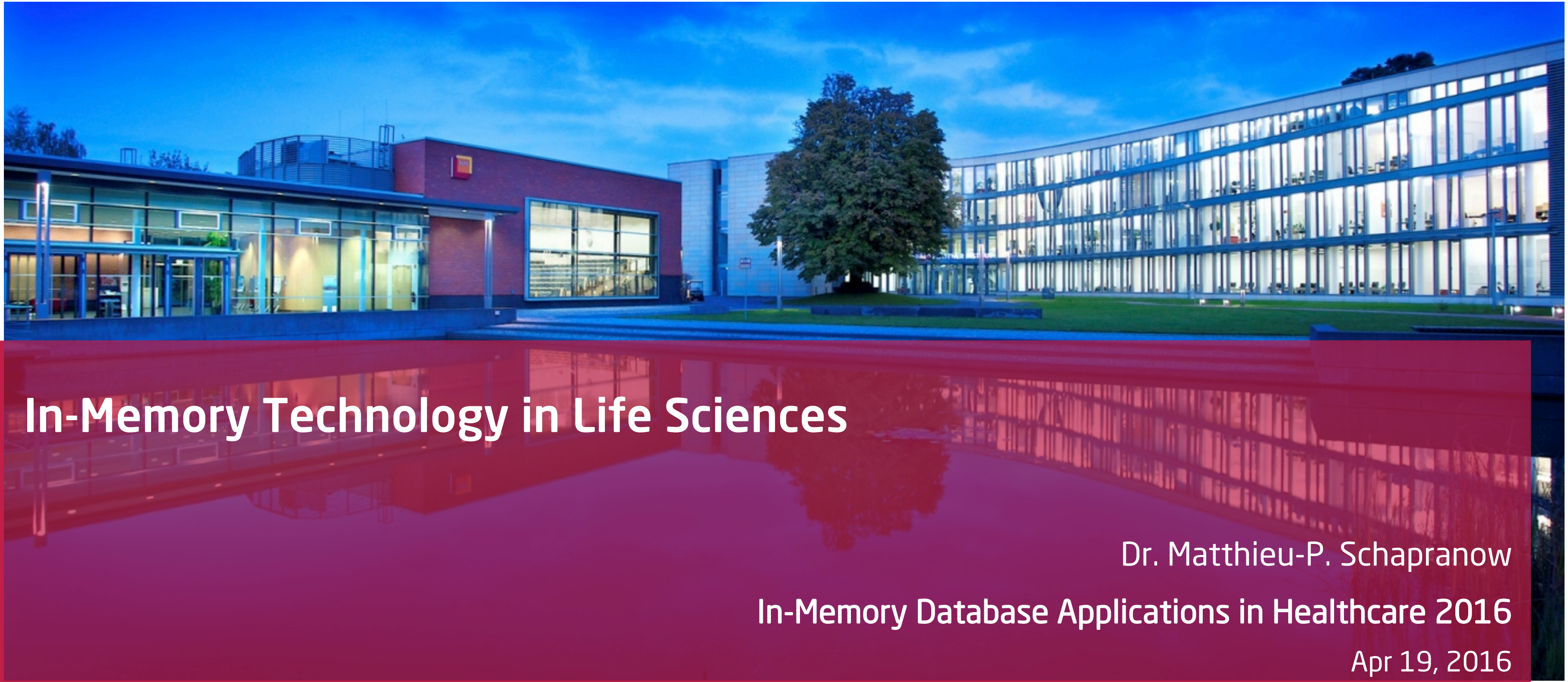
Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



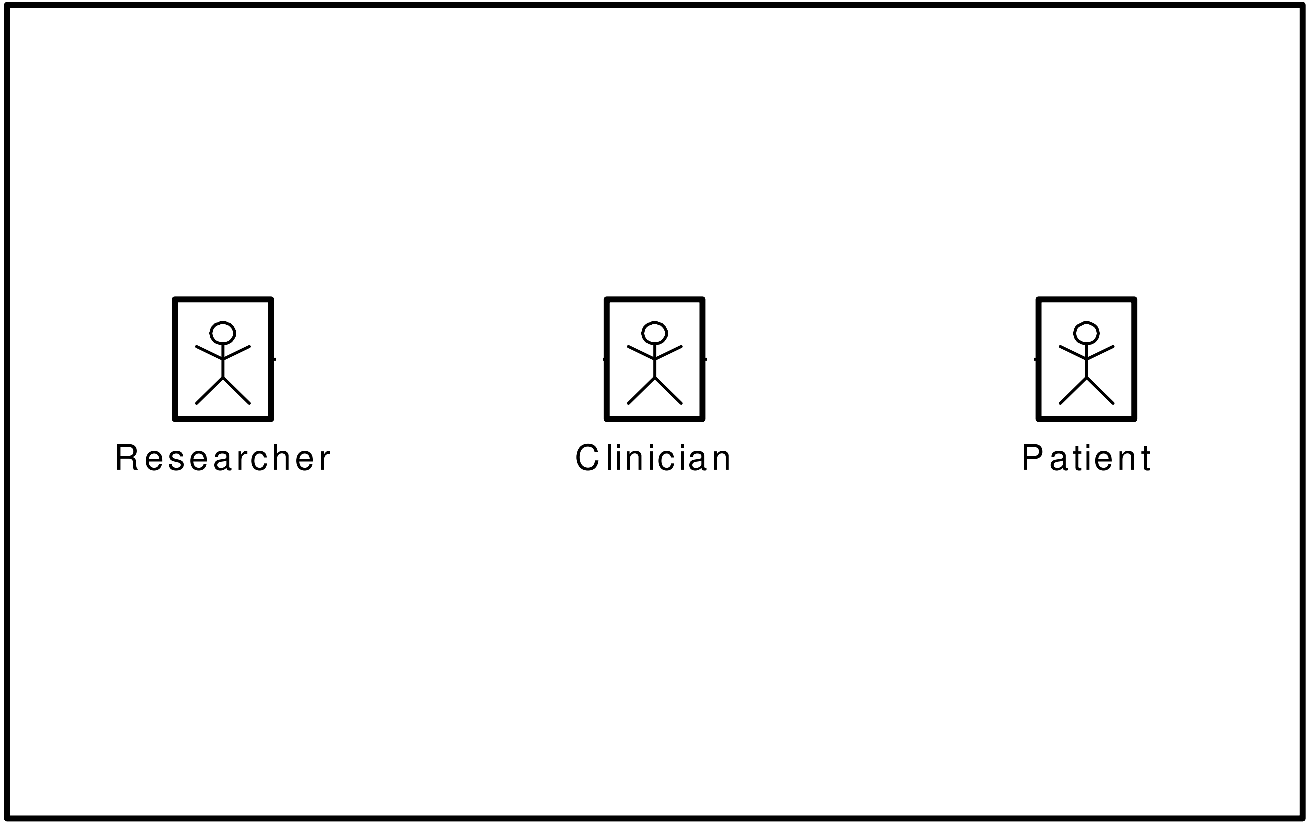
In-Memory Technology in Life Sciences

Dr. Matthieu-P. Schapranow

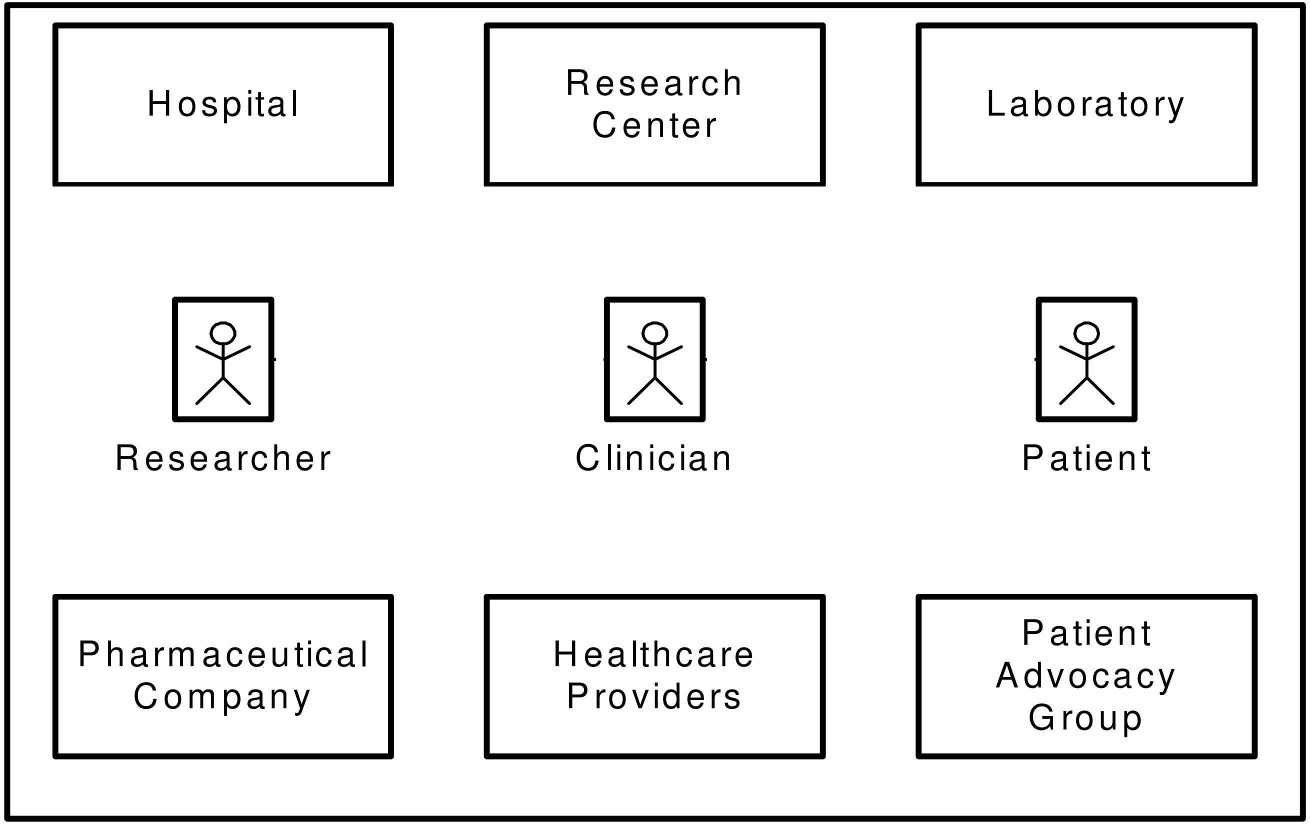
In-Memory Database Applications in Healthcare 2016

Apr 19, 2016

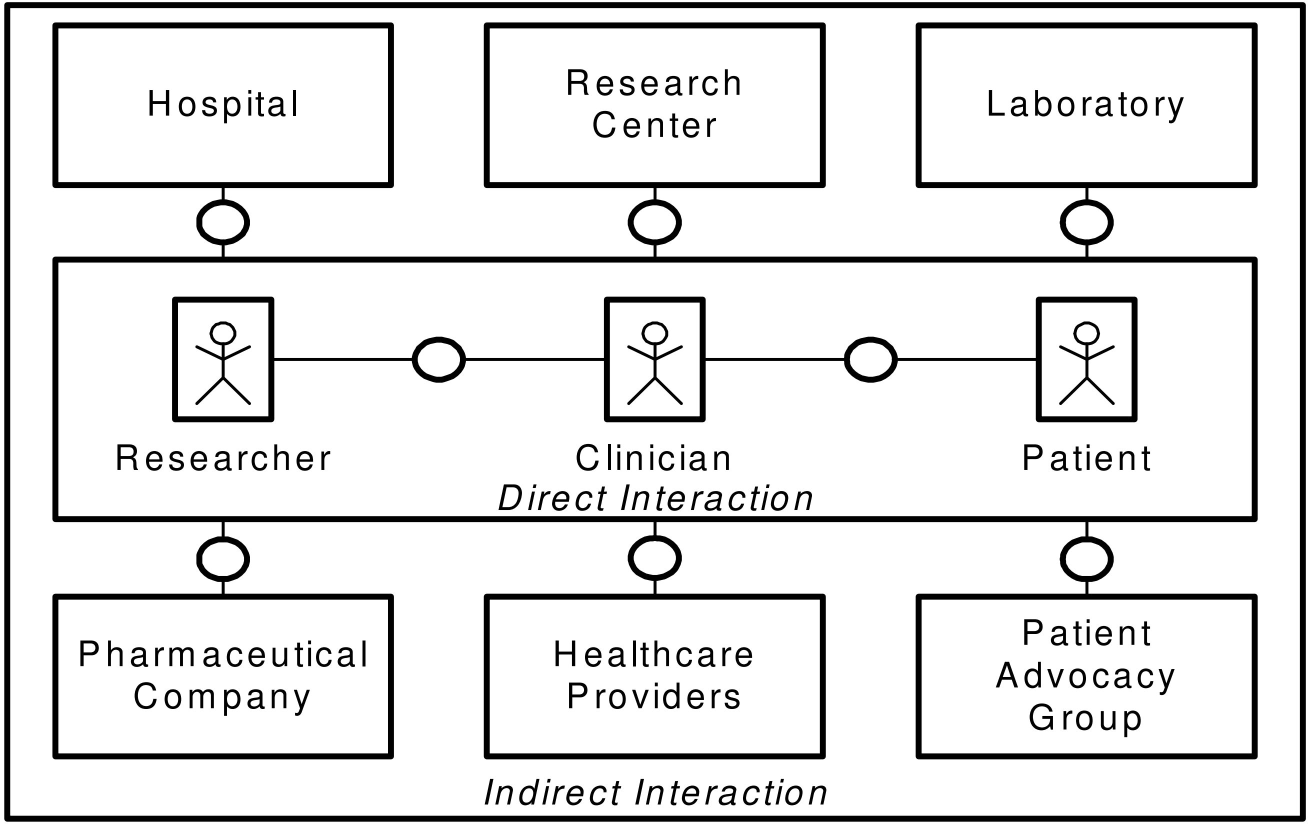
Intelligent Healthcare Networks in the 21st Century?



Intelligent Healthcare Networks in the 21st Century?

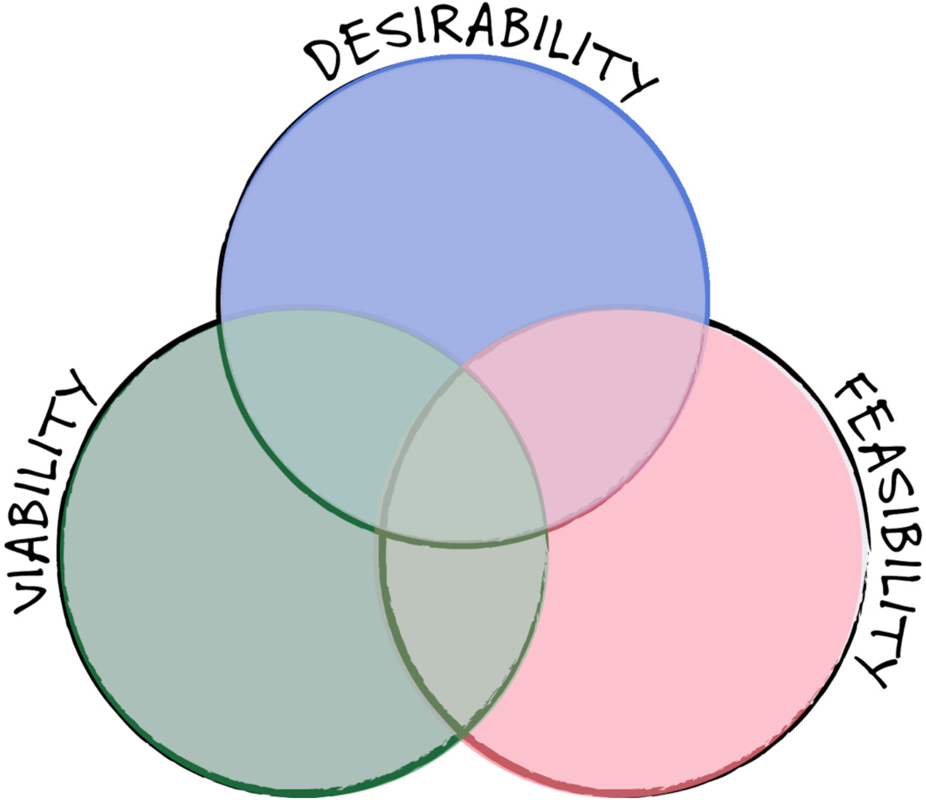


Intelligent Healthcare Networks in the 21st Century!



Our Methodology
Design Thinking

Hasso Plattner Institut



Our Methodology
Design Thinking

Desirability

- Portfolio of integrated services for clinicians, researchers, and patients
- Include latest treatment option, e.g. most effective therapies

Viability

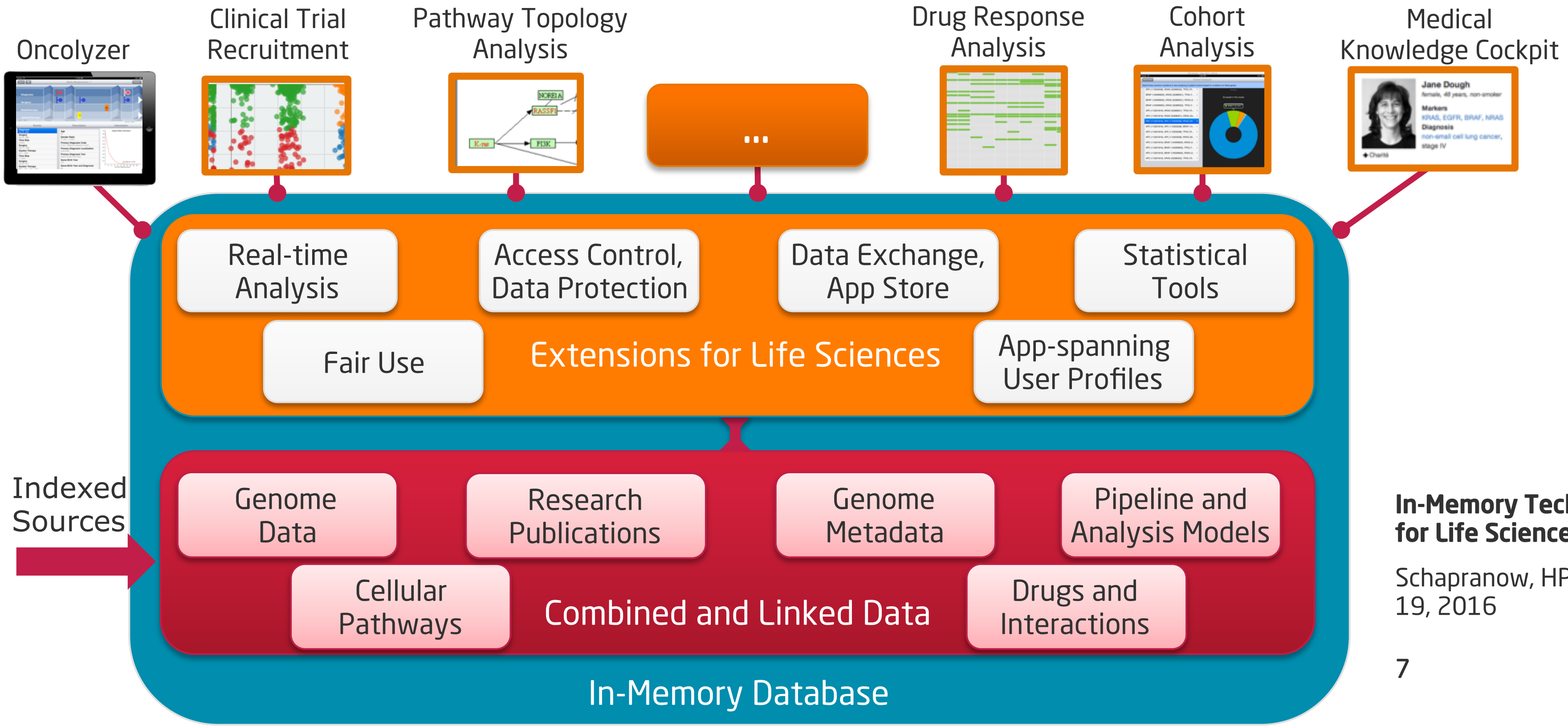
- Enable precision medicine also in far-off regions and developing countries
- Involve word-wide experts (cost-saving)
- Combine latest international data (publications, annotations, genome data)

Feasibility

- HiSeq 2500 enables high-coverage whole genome sequencing in 20h
- IMDB enables allele frequency determination of 12B records within <1s
- Cloud-based data processing services reduce TCO

we.analyzegenomes.com

Real-time Analysis of Big Medical Data



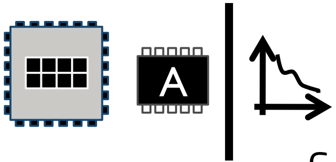
In-Memory Database Technology
Use Case: Analysis of Genomic Data

Analysis of Genomic Data

	Alignment and Variant Calling	Analysis of World-wide Annotations
Bound To	CPU Performance	Memory Capacity
Duration	Hours – Days	Weeks
HPI	Minutes	Real-time
In-Memory Technology	Multi-Core	Partitioning & Compression

In-Memory Data Management Overview

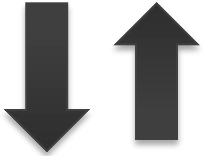
Advances in Hardware



Multi-core architecture (6 x 12 core CPU per blade)

Parallel scaling across blades

1 blade ≈50k USD = 1 enterprise class server



64 bit address space - 4 TB in current server boards

4 MB/ms/core data throughput

Cost-performance ratio rapidly declining

Advances in Software

Row and Column Store

Insert Only

Compression

Partitioning

Parallelization

Active & Passive Data Stores

In-Memory Database Technology
Hardware Characteristics at HPI FSOC Lab

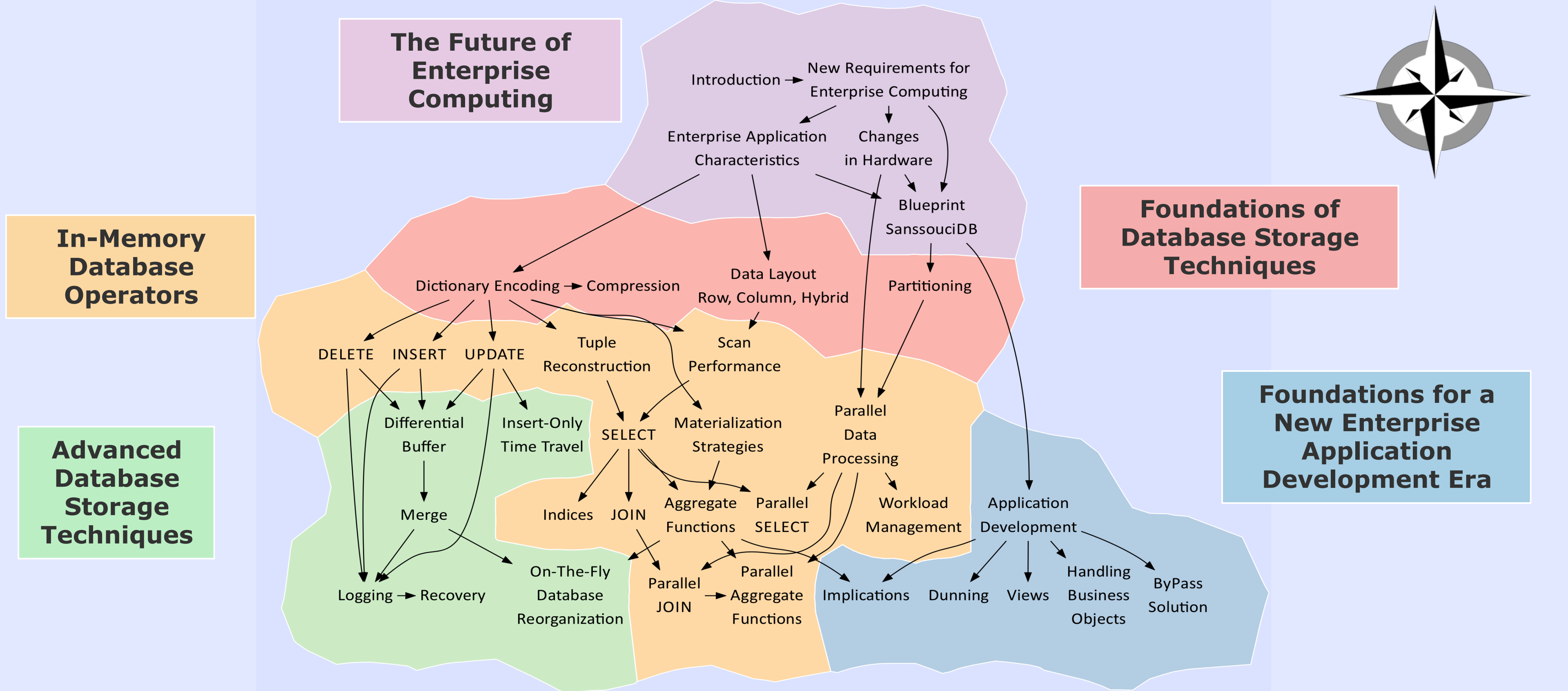
- **1,000 core cluster at Hasso Plattner Institute with 25 TB main memory**
- 25 nodes, each consists of:
  - 40 cores
  - 1 TB main memory
  - Intel® Xeon® E7- 4870
  - 2.40GHz
  - 30 MB Cache

Our Technology
In-Memory Database Technology

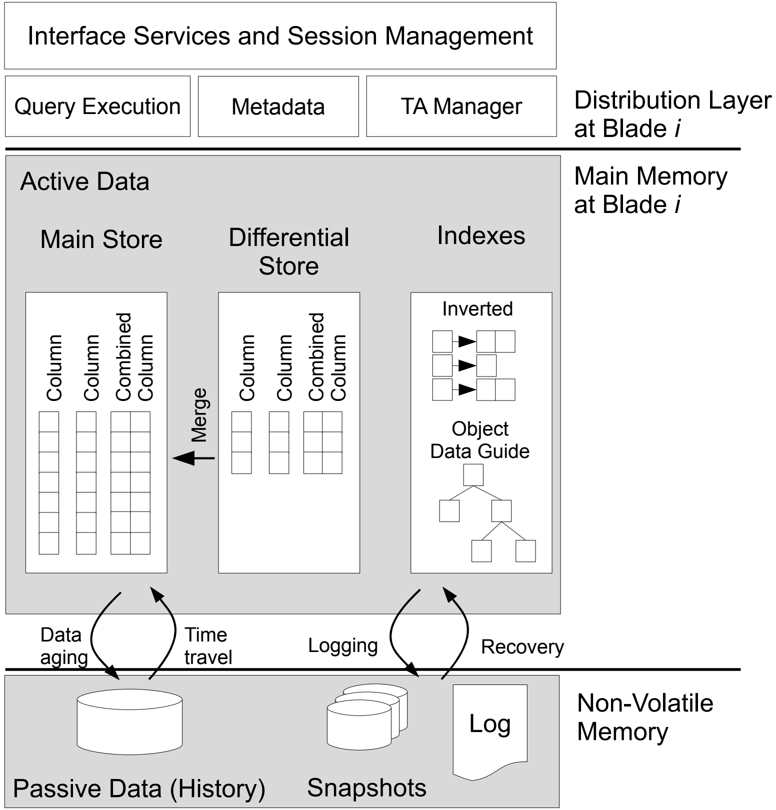
Hasso Plattner Institut

- Combined column and row store
- Map/Reduce
- Single and multi-tenancy
- Lightweight compression
- Insert only for time travel
- Real-time replication
- Working on integers
- SQL interface on columns and rows
- Active/passive data store
- Minimal projections
- Group key
- Reduction of software layers
- Dynamic multi-threading
- Bulk load of data
- Object-relational mapping
- Text retrieval and extraction engine
- No aggregate tables
- Data partitioning
- Any attribute as index
- No disk
- On-the-fly extensibility
- Analytics on historical data
- Multi-core/ parallelization

Learning Map of openHPI Course In-Memory Data Management



SanssouciDB: An In-Memory Database for Enterprise Applications



Lightweight Compression

- Typical compression factor of 10:1 for enterprise software
- In financial applications up to 50:1

Hasso Plattner Institut

- Main memory access is the new bottleneck
- Lightweight compression can reduce this bottleneck, i.e.
  - Lossless
  - Improved usage of data bus capacity
  - Work directly on compressed data

Table

RecId 1	091487	Colon	C18.0
RecId 2	357982	Larynx	C32.0
RecId 3	123489	Lip	C00.9
RecId 4	998711	Colon	C18.0
RecId 5	215678	Rectum	C20.0
RecId 6	647912	Rectum	C20.0
RecId 7	167898	Mama	C50.9
RecId 8	646470	Colon	C18.0
…	…	…	…

Attribute Vector

RecId	ValueId
1	C18.0
2	C32.0
3	C00.9
4	C18.0
5	C20.0
6	C20.0
7	C50.9
8	C18.0

Data Dictionary

ValueId	Value
1	Larynx
2	Lip
3	Rectum
4	Colon
5	Mama

Inverted Index

ValueId	RecIdList
1	2
2	3
3	5,6
4	1,4,8
5	7

Combined Column and Row Store

Row Stores	Column Stores
Designed for operative workload, e.g.	Designed for analytical work, e.g.
• Create and maintain meta data for laboratory tests	• Evaluate the number of positive test results
• Access a complete record of a clinical trial or experiment series	• Identification of correlations or test candidates

- In-Memory Technology combines both stores
  - Increased performance for analytical work
  - Operative performance remains interactively

Insert-Only / Append-Only

- Traditional databases allow four data operations:
  - INSERT, SELECT and
  - DELETE, UPDATE (destructive)

- Insert-only database tables
  - INSERT, SELECT performed, DELETE, UPDATE are built on them
  - Maintain complete history, e.g. bookkeeping systems
  - Enable time travelling, e.g. to
    - Trace changes and reconstruct medical decisions
    - Document complete history of changes in therapies, dosages, etc.
    - Enable statistical observations of blood pressure, heart rate, etc.

Data Partitioning

Horizontal Partitioning	Vertical Partitioning
Cut long tables into shorter segments	Split off selected columns to individual resources
Example: Grouping of samples belonging to same experiment, patients of the same station, etc.	Example: Separation of personalized data from experiment data, research vs. clinical data

- IMDB supports both partitioning approaches
- Data Partitioning is the basis for
  - Parallel execution of database queries
  - Implementation of data aging and data retention management

Multi-core and Parallelization

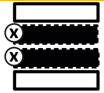
- Modern server systems consist of *x* CPUs, e.g. 6
- Each CPU consists of *y* CPU cores, e.g. 12
- Consider each of the *x***y* CPU core as individual **workers, e.g. 6x12 = 72**
- Each worker can perform one task at the same time in parallel

- Full table scan of database table w/ 1M entries results in 1/x*1/y processing time when traversing in parallel
  - Reduced response time
  - No need for pre-aggregated totals and redundant data
  - Improved usage of hardware
  - Instant analysis of data

Active and Passive Data Store

Active Data	Passive Data
Accessed and updated frequently, e.g.	Used for analytical & statistical purposes only, e.g.
• Most recent experiment results, e.g. last two weeks	• Samples that were processed 5 years ago
• Samples that have not been processed, yet	• Meta data about seeds that are not longer produced

- Passive data can be stored on slower storages
  - Reduces main memory demands
  - Improves performance active data



Reduction of Application Layers

- Layers are introduced to abstract software complexity
- Each layer offers complete functionality, e.g. meta data of samples

- Less layers result in
  - More specific code only
  - Improved code maintainability
  - Reduced resource demands
  - Improved performance of applications due to eliminating obsolete processing

Keep in contact with us!

Hasso
Plattner
Institut

Dr. Matthieu-P. Schapranow
schapranow@hpi.de
http://we.analyzegenomes.com/

Hasso Plattner Institute
Enterprise Platform & Integration Concepts (EPIC)
Program Manager E-Health
Dr. Matthieu-P. Schapranow
August-Bebel-Str. 88
14482 Potsdam, Germany