



# SCADS RINGVORLESUNG FÜR BIG DATA

KOORDINATOREN:

- PROF. DR. S. GUMHOLD, TU DRESDEN
- PROF. DR. E. RAHM, UNIV. LEIPZIG

[www.scads.de/de/lehre/scads-ringvorlesung](http://www.scads.de/de/lehre/scads-ringvorlesung)



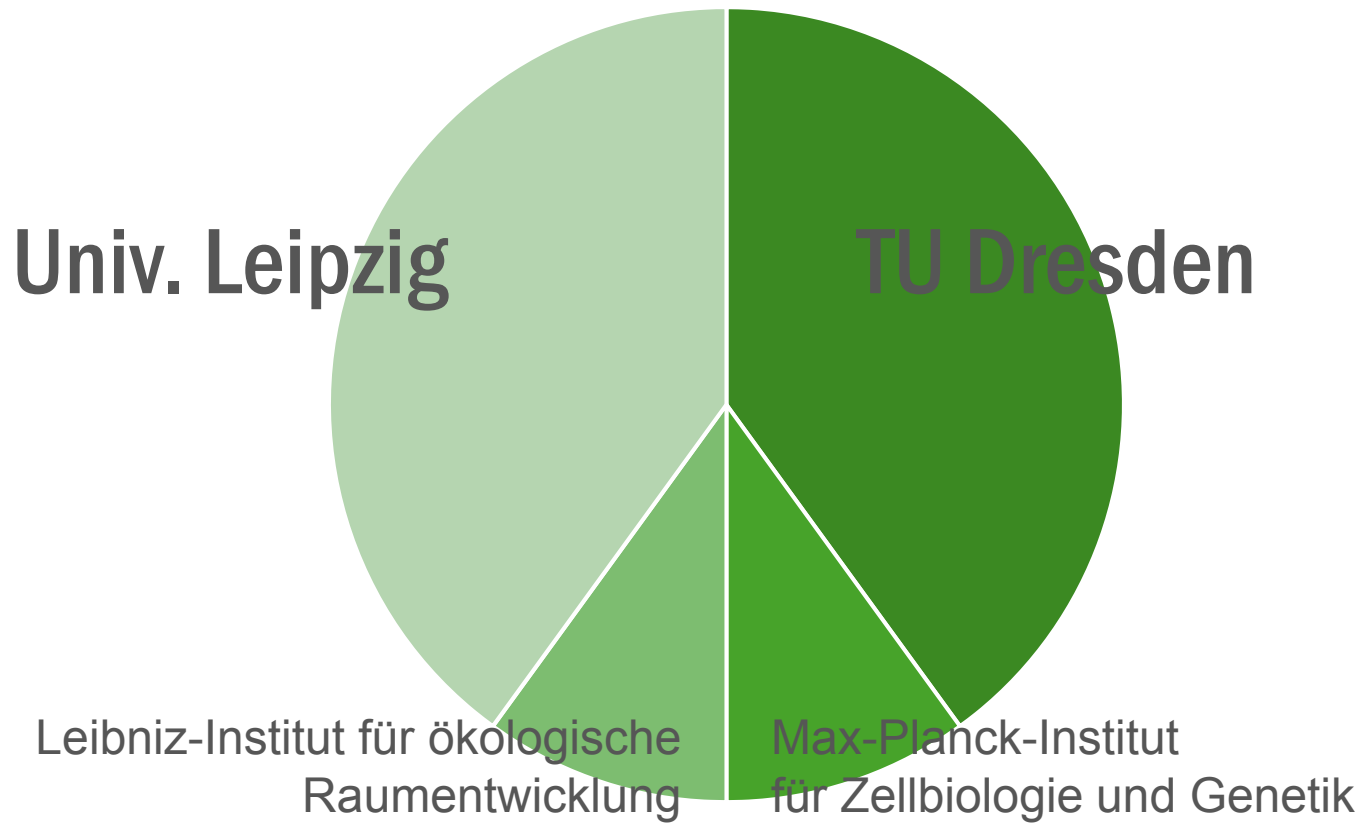
Zwei deutsche Kompetenzzentren für Big Data seit Okt. 2014 (BMBF-Wettbewerb)

- ScaDS Dresden/Leipzig
- Berlin Big Data Center (BBDC)

**ScaDS Dresden/Leipzig (Competence Center for Scalable Data Services and Solutions Dresden/Leipzig)**

- wissenschaftliche Koordinatoren: Nagel (TUD), Rahm (UL)
- Laufzeit: zunächst 4 Jahre (bis Sep. 2018)
- Option zur Verlängerung um weitere 3 Jahre







- Avantgarde-Labs GmbH
- Data Virtuality GmbH
- E-Commerce Genossenschaft e. G., Leipzig
- European Centre for Emerging Materials and Processes (ECEMP), Dresden
- Fraunhofer-Institut für Verkehrs- und Infrastruktursysteme (IVI), Dresden
- Fraunhofer-Institut für Werkstoff- und Strahltechnik (IWS), Dresden
- GISA GmbH, Halle
- Helmholtz-Zentrum Dresden - Rossendorf
- Hochschule für Telekommunikation Leipzig (HfTL)
- Institut für Angewandte Informatik e. V. (InfAI), Leipzig
- Landesamt für Umwelt, Landwirtschaft und Geologie
- Netzwerk Logistik Leipzig-Halle e. V.
- Sächsische Landesbibliothek – Staats- und Universitätsbibliothek (SLUB) Dresden
- Scionics Computer Innovation GmbH
- Technische Universität Chemnitz
- Umweltforschungszentrum (UFZ) Leipzig
- Universitätsklinikum Carl Gustav Carus, Dresden



## GROBSTRUKTUR DES ZENTRUMS

Lebenswissenschaften

Werkstoff- und Ingenieurwissenschaften

Umwelt- /Verkehrswissenschaften

Digital Humanities

Business Data

Service-  
zentrum

Big Data Life Cycle Management und Workflows

Datenqualität /  
Datenintegration

Wissensextraktion

Visuelle  
Analyse

Effiziente Big Data Architekturen



## Zielsetzung

- Überblick über aktuelle Forschung und Anwendungen zu Big Data
- Vorstellung der ScaDS-Themen durch jeweilige Professoren/PIs

## Modus

- 6 Termine mit je 2 Vorträgen a ca 1 h
- jeweils donnerstags ab 15 Uhr
- abwechselnd in Leipzig (HS8) und Dresden (Willersbau W317)

## Zielgruppen

- Studierende der beiden Universitäten, u.a. in Master- und Bachelorstudiengängen der Informatik
- Doktoranden und Forscher
- alle Interessenten



## Block 1: 27. April, Leipzig

- Prof. Rahm: Datenintegration und Graph-Analysen für Big Data
- Prof. Scheuermann: Merkmalsbasierte visuelle Analyse großer wissenschaftlicher Daten

## Block 2: 11. Mai, Dresden

- Prof. Sbalzarini: The PPML language for distributed scalable processing enables real-time segmentation of large image data
- Prof. Lehner: Next-Generation Hardware for Data Management – more a Blessing than a Curse?

## Block 3: 18. Mai, Leipzig

- Prof. Heyer: Big Data in den Digital Humanities?
- Prof. Stadler: Genome Annotation in the Age of Big Data





# Modern Hardware – All over the place...



## Next-Gen Hardware for Data Management

Wolfgang Lehner – ScaDS Ringvorlesung – 11.5.2017 - Willersbau W317

### Utilization Wall: Dark Silicon's Effect on Multicore Scaling

Spectrum of tradeoffs between # of cores and frequency

Example: 65 nm → 32 nm (S=2)

2x4 cores @ 1.8 GHz (8 cores dark, 8 dim)

### Transaction Logging Unleashed with NVRAM\*

Tianzheng Wang, Ryan Johnson  
University of Toronto  
(tzwang, ryan.johnson)@cs.toronto.edu

PostgreSQL users are first buffered against parallel hardware, else a bottleneck: prior with hyperparallel because the 85% of the CPU

### Fast Updates on Read-Optimized Databases Using Multi-Core CPUs

Jens Krueger<sup>1</sup>, Changkyu Kim<sup>1</sup>, Martin Grund<sup>1</sup>, Nadathur Satish<sup>1</sup>, David Schwalb<sup>1</sup>, Jatin Chhugani<sup>1</sup>, Hasso Plattner<sup>1</sup>, Pradeep Dubey<sup>1</sup>, Alexander Zeier<sup>1</sup>

<sup>1</sup>Hasso-Plattner-Institute, Potsdam, Germany  
Contact: jens.krueger@hpi.uni-potsdam.de

<sup>2</sup>Parallel Computing Lab, Intel Corporation  
Contact: changkyu.kim@intel.com

### Data-Oriented Transaction Execution

Ippokratis Pandis<sup>1,2</sup>, Ryan Johnson<sup>1,2</sup>  
ipandis@ece.cmu.edu

### SGI Scales Up HANA On UV NUMA

June 3, 2011. By Timothy Prickett Morgan

4 cores @ 1.8 GHz

65 nm

### Starting Concurrent

Xiaoyu Chen  
xyx@cs.cmu.edu

Andrew Pavlo  
Carnegie Mellon University  
pavlo@cs.cmu.edu

Srinivas Devadas  
MIT CSAIL  
devadas@csail.mit.edu

Michael Stonebraker  
MIT CSAIL  
stonebraker@csail.mit.edu

### Subsystem for Modern Hardware

David Lomet  
Microsoft Research  
One Microsoft Way  
Redmond, WA 98052  
lomet@microsoft.com

Sudipta Sengupta  
Microsoft Research  
One Microsoft Way  
Redmond, WA 98052  
sudipta@microsoft.com

### Staring Concurrent

Abstract: Computer architectures are moving towards an era dominated by many-core machines with dozens or even hundreds of cores on a single chip. This unprecedented level of on-chip parallelism introduces a new dimension to scalability that current database management systems (DBMSs) were not designed for. In particular, as the number of cores increases, the problem of concurrency control becomes extremely challenging. With hundreds of threads running in parallel, the complexity of coordinating competing accesses to data will likely diminish the gains from increased core counts.

To better understand just how unprepared current DBMSs are for future CPU architectures, we performed an evaluation of concurrency control for on-line transaction processing (OLTP) workloads on many-core chips. We implemented seven concurrency control algorithms on a state-of-the-art DBMS and used a synthetic workload that intentionally forces contention to occur on all cores.

Abstract: LLAMA is a subsystem designed for new hardware environments that supports an API for page-oriented access methods, providing both cache and storage management. Caching (CL) and storage (SL) layers use a common mapping table that separates a page's logical and physical location. CL supports data updates and management updates (e.g., for index re-organization) via latch-free compare-and-swap atomic state changes on its mapping table. SL uses the same mapping table to cope with page location changes produced by log structuring on every page flush. To demonstrate LLAMA's suitability, we tailored our latch-free Bw-tree implementation to use LLAMA. The Bw-tree is a B-tree style index. Layered on LLAMA, it has higher performance and scalability using real workloads compared with BerkeleyDB's B-tree, which is known for good performance.





### Block 4: 1. Juni, Dresden

- Prof. Nagel: Big Data und HPC - zwei Welten oder eine gemeinsame Zukunft?
- Dr. Bussmann: Big Data in Photon Science: Why we do everything once

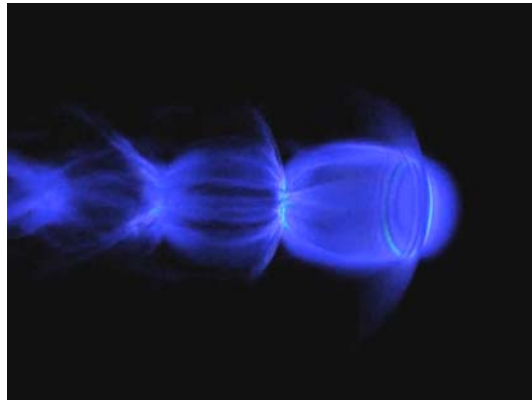
### Block 5: 22. Juni, Leipzig

- Prof. Bogdan: Verbesserung der Sicherheit von Virtuellen Maschinen für Big Data Architekturen
- Prof. Franczyk: Prozesse treffen Big Data – Verbindung zwischen Data Science und Process Science

### Block 6: 29. Juni, Dresden

- Prof. Gumhold: Scalable Visualization
- Prof. Dachsel: Multimodal Exploration of Large Data Sets



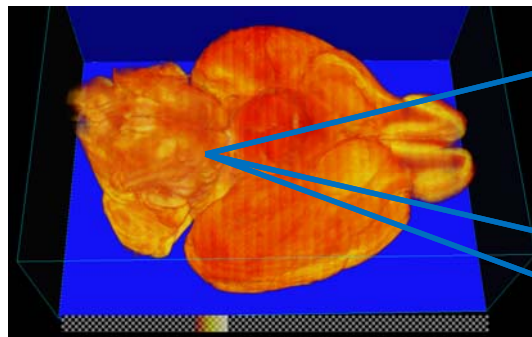


laser-plasma accelerator

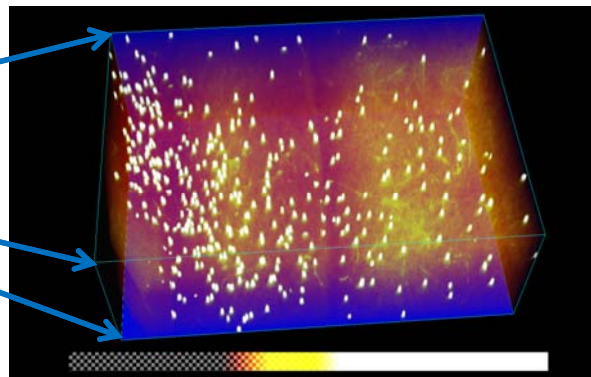


Computergraphik  
und Visualisierung

**Scalable Visualization**  
Interactive Rendering of  
huge volume data



fruit fly embryo

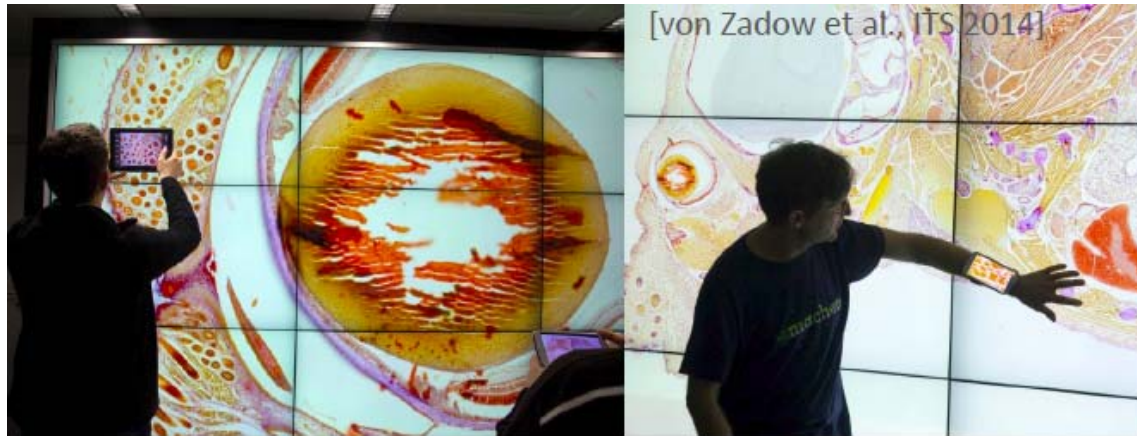


zoom into cellular details

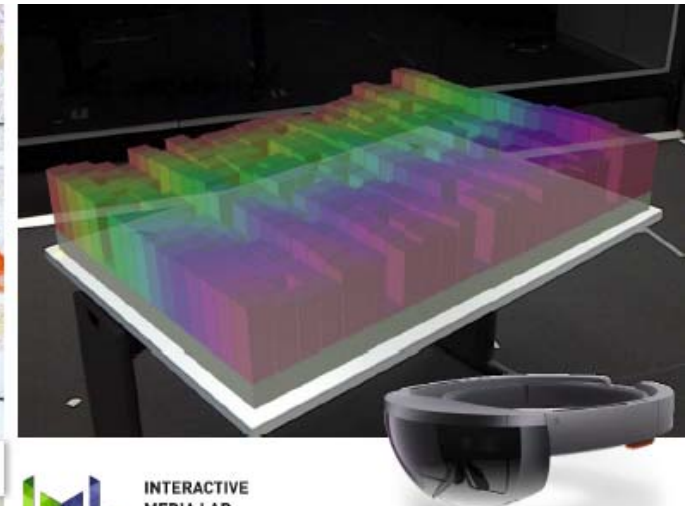


visible human



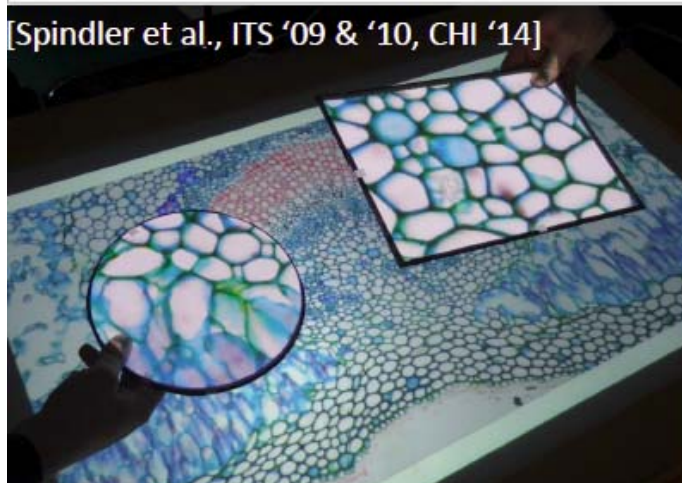


[von Zadow et al., ITS 2014]



ScaDS 2017: Multimodal  
Exploration of Large Data Sets

Prof. Dr.-Ing. Raimund Dachzelt



[Spindler et al., ITS '09 & '10, CHI '14]



[Büschel et al., WS IA 2016]



[Kister et al., ITS 15]



## Belegung in folgenden Modulen (2/2/0)

- Bachelor Informatik und Medieninformatik: INF-B-510, INF-B-520, INF-B-530, INF-B-540
- Master Medieninformatik: INF-BAS7, INF-VERT7
- Master und Diplom Informatik: INF-BAS7, INF-VMI-8

## Leistungserbringung

- Teilnahme an den sechs Vorlesungsterminen
- **schriftliche Ausarbeitung** (ca 15 Seiten) zu drei der 12 Themen, davon mind. zwei von Dresdner Referenten **oder Bearbeitung einer praktischen Aufgabe** zu einem Dresdner Thema



Belegung als Modul „Aktuelle Trends der Informatik“ (5 credits)

- fakultätsinterne Schlüsselqualifikation für Bachelor und Master Informatik

ggf. als Teilleistung für Vertiefungsmodul „Anwendungsbezogene Datenbankkonzepte“

- falls praktische Leistung zum Datenbankthema (Prof. Rahm)

### Leistungserbringung

- Teilnahme an den sechs Vorlesungsterminen
- **schriftliche Ausarbeitung** (ca 15 Seiten) zu drei der 12 Themen, davon mind. zwei von Leipziger Referenten **oder Bearbeitung einer praktischen Aufgabe** zu einem Leipziger Thema
- Themenvergabe der prakt. Aufgaben erfolgt im Anschluss an heutige Vorträge