Tim Kaldewey

Curriculum Vitae

505 Governors Ct.
Philadelphia, PA 19146
http://www.kaldewey.com

PHONE: (650) 799-1189
E-MAIL: tim@kaldewey.com

EDUCATION

Ph. D., Computer Science, University of California, Santa Cruz, March 2010
Thesis: *Predictable High-Performance Data Management – Leveraging System Resource Characteristics to Improve Performance and Predictability*
Advisor: Prof. Scott Brandt
Committee members: Prof. Andrea Di Blas, Prof. Charlie McDowell, Prof. Richard Hughey, Oracle Architect Eric Sedlar

M. S., Computer Science, University of California, Santa Cruz, March 2008
Thesis: *Virtualizing Storage Performance*
Advisor: Prof. Scott Brandt

M. S., Network Engineering, Institute Eurecom, Sophia Antipolis, FRANCE & University of Technology Darmstadt, GERMANY, September 2004
Thesis: *Location Tracking Services, Business Case: Incident command post for emergency response*
Advisor: Prof. Pascal Felber

HONORS

IBM awards	2015, 2014, 2013
Best paper awards	SIGMOD 2010, RTAS 2008
Oracle Fellowship	2010, 2009, 2008
University of California Regent's Fellowship	2006, 2005
Eurecom Scholarship	2003

INTERESTS

Performance/efficiency optimization of AI workloads

Parallel algorithms optimized for parallel architectures

System resource management to guarantee application performance

High-performance and large-scale data management on parallel systems

EMPLOYMENT HISTORY

May 2014–present **Performance Architect**, IBM Watson Group (R&D), New York, New York
Lead performance/efficiency optimization efforts on Watson offerings through the hardware/software stack across the Watson organization. [publications C16–20, D5, J5]

October 2011–present **Assistant Adjunct Professor**, Computer and Information Science Department, University of Pennsylvania, Philadelphia.

Lectures on selected topics from systems, architecture, and parallel programming.

November 2010–Apr 2014 **Research Staff Member**, IBM Almaden Research – Database Technologies, San Jose, California

Ongoing research on high-performance data management on emerging parallel architectures, large-scale distributed systems, and high-performance storage systems. [publications C12–15, D1–4, B1, P6]

April 2008–November 2010 **Senior Researcher**, Oracle Corporation, Server Technologies – Special Projects, Redwood Shores, California

Research on high-performance data management on parallel systems. Developed novel parallel algorithms for large-scale data management on emerging parallel architectures, e.g. multi-core CPUs, vector units, graphics processors. Prototype implementations demonstrated performance increases of more than an order of magnitude. [publications C8–11, J2–4]

Technology transfers to product groups, e.g. parallel string comparison, parallel memory copy, parallel join.

January 2006–March 2008 **Graduate Student Researcher**, Real Time Lab – Storage Systems Research Center (SSRC), Department of Computer Science, University of California Santa Cruz.

Research on virtualizing storage performance within a distributed system including disk scheduling, caching and networking. [C4, C5, C7, J1]

Investigated virtualization of storage devices using real-time scheduling techniques. Developed a disk scheduler that provides perfect isolation of workloads and performance guarantees in shared storages system. Implemented prototype as Linux kernel module. [C6]

Designed an extension for distributed filesystems to accelerate access to *in-flight* data, i.e.data currently in transit in a storage system.

Added support for firm real-time processing to an integrated real-time scheduler for the Linux 2.6 kernel. Firm real-time workloads have no value when deadlines are missed but parts may be skipped, e.g. B-frames in a video stream. [C2,C3]

June 2007–September 2007 **Research Intern**, Oracle Corporation, Server Technologies – Special Projects, Redwood City, California

Research on relational database functions for parallel systems. Developed parallel algorithms for core database functions including index search, post-filter, and join. Implemented and evaluated prototypes running on graphics cards, multi-core CPUs, and vector units. [C9]

June 2006–September 2006

Research Intern, IBM Almaden Research Center, San Jose, California

Research on Storage QoS. Developed and evaluated a Real-Time disk scheduler that implements a virtual disk, providing perfect isolation between multiple workloads on a shared disk. The prototype was implemented as a Linux kernel device driver and uses disk (time) utilization as a metric of storage performance. [C6]

March 2004–August 2005

Research Engineer, SAP Research, Palo Alto, California

Designed and developed **SAP Location Service Engine**. Basic functionality included display of real-time location(s) inside a building, indoor navigation, restricted area access alerts, and proximity search. Prototype applications comprised an incident command post for emergency response scenarios and asset tracking. The system is independent from specific sensing technologies and interfaces with the SAP AutoID node infrastructure. [P1-P5, W1]

September 2000–March 2003

Fleet Management - Technical Assistant, Lufthansa Technik AG, Aircraft Systems Engineering Airbus Fleet, Frankfurt Airport, GERMANY

Managed maintenance schedule for Lufthansa's Airbus fleet in SAP R/3, ensuring on-time compliance with aviation authority requirements, manufacturer specifications, and Lufthansa safety standards. This position required security clearance for unrestricted hangar & airport access to verify compliance with ordered changes.

September 1997–December 2000

Network Specialist, Institute of Management and Logistics, University of Technology Darmstadt, GERMANY

Managed Novell Netware & Windows network. Responsible for roll out of Win NT 4.0, W2K, etc. Designed and implemented security policies for application and file server access.

December 1996–July 1997

Network Specialist, IT services, Software AG systems, Darmstadt, GERMANY

Managed Novell Netware & Windows network. Performed troubleshooting of network issues, analyzed network performance and suggested improvements of network infrastructure. Migration from a Token Ring to an Ethernet network.

January 1996–September 1996

Network Specialist, IT department, Anneliese Zementwerke AG Ennigerloh, GERMANY

Managed Novell Netware, Windows & SAP R3 servers. Conducted network performance analysis, designed network infrastructure for subsidiaries, set up WAN connectivity to corporate data center, designed and implemented a security policies for Novell and Windows networks.

TEACHING EXPERIENCE

October 2016 **Lecture** in CIS 565 – GPU Programming and Architecture, Department of Computer and Information Science, University of Pennsylvania.
“*Watson from a performance perspective – Enhancing, scaling, and accelerating human expertise* ”

October 2015 **Lecture** in CIS 565 – GPU Programming and Architecture, Department of Computer and Information Science, University of Pennsylvania.
“*Accelerating Watson Watson Workloads – Enhancing, scaling, and accelerating human expertise* ”

October 2014 **Lecture** in CIS 565 – GPU Programming and Architecture, Department of Computer and Information Science, University of Pennsylvania.
“*Programming GPUs for database applications – outsourcing index search operations*”

April 2014 **Lecture** in CIS 565 – GPU Programming and Architecture, Department of Computer and Information Science, University of Pennsylvania.
“*Programming GPUs for non-graphics workloads - from General Purpose GPU (GPGPU) to GPU compute*”

December 2013 **Lecture** in CIT 593 – Computer Systems I, Department of Computer and Information Science, University of Pennsylvania.
“*x86 Architecture*”

October 2013 **Lecture** in CIS 565 – GPU Programming and Architecture, Department of Computer and Information Science, University of Pennsylvania.
“*GPU Database Workloads*”

September 2013 **Lecture** in CIS 565 – GPU Programming and Architecture, Department of Computer and Information Science, University of Pennsylvania.
“*Large-Scale Data Management on GPU*”

November 2012 **Lecture** in CIT 593 – Computer Systems I, Department of Computer and Information Science, University of Pennsylvania.
“*x86 Architecture*”

October 2012 **Lecture** in CIS 565 – GPU Programming and Architecture, Department of Computer and Information Science, University of Pennsylvania.
“*GPU search*”

February 2012 **Lecture** in CIS 565 – GPU Programming and Architecture, Department of Computer and Information Science, University of Pennsylvania.
“*Large-Scale GPU Programming*”

November 2011 **Lecture** in CIS 565 – GPU Programming and Architecture, Department of Computer and Information Science, University of Pennsylvania.
“*Programming GPUs for database applications*”

November 2011 **Lecture** in CIT 593 – Computer Systems I, Department of Computer and Information Science, University of Pennsylvania.
“*x86 Architecture*”

August 2010

Workshop on GPU programming, Oracle Labs.

Organized workshop including guest speaker for lecture on GPU architecture. Gave lectures on "*Introduction to GPU programming*" and "*Optimizations for data-intensive applications*".

February 2010

Workshop on Architecture Aware Programming, Oracle Special Projects & SunLabs.

Co-organized workshop and presentation on "*optimizing memory access patterns and performance*".

March 2009

Guest Lecture in CMPE 220 – Advanced Parallel Programming, Department of Computer Engineering, University of California Santa Cruz.

"*SSE Vector programming on x86 architectures*"

March 2008

Guest Lecture in CMPE 220 – Advanced Parallel Programming, Department of Computer Engineering, University of California Santa Cruz.

"*SSE Vector programming on x86 architectures*"

September 2005–December 2005

Teaching Assistant for CMPS 111 – Introduction to Operating Systems (Upper division undergraduate class), Department of Computer Science, University of California Santa Cruz.

Gave lectures on programming and homework assignments. Supervised students during lab hours. Graded homework and programming assignments. Programming assignments comprised implementing a shell, a CPU scheduler, multitasking, and a file system.

PUBLICATIONS

Conference & Workshop Papers

C20. T. Kaldewey, D. K. Tam, “Optimizing Efficiency of Deep Learning Workloads through GPU Virtualization”, *GPU Technology Conference* (**GTC’17**), San Jose, California May 8–11, 2017.

C19. T. Kaldewey, D. Wendt, “Accelerating Document Retrieval and Ranking for Cognitive Applications”, *GPU Technology Conference* (**GTC’17**), San Jose, California May 8–11, 2017.

C18. E. Sitaridi, R. Mueller, T. Kaldewey, G. Lohman, K. Ross “Massively-Parallel Lossless Data Decompression”, *45th Annual International Conference on Parallel Processing* (**ICPP’16**), Philadelphia, Pennsylvania August 16–19, 2016.

C17. P. Haggar, M. Gschwind, and T. Kaldewey “Towards an optimized deep learning infrastructure on OpenPOWER”, *1st IBM Deep Learning Workshop*, Yorktown Heights, New York January 29, 2016.

C16. E. Sitaridi, R. Mueller, T. Kaldewey, “Parallel Lossless Compression using GPUs”, *GPU Technology Conference* (**GTC’14**), San Jose, California March 24–27, 2014.

C15. R. Gandhi, A. Gupta, A. Povzner, W. Belluomini, and T. Kaldewey “Mercury: Bringing Efficiency to Key-value Stores”, *6th International Systems and Storage Conference* (**Systor’13**), Haifa, Israel, June 30–July 2, 2013.

C14. T. Kaldewey, R. Mueller “Let your GPU do the heavy lifting in your data warehouse.”, *GPU Technology Conference* (**GTC’13**), San Jose, California March 18–21, 2013.

C13. T. Kaldewey, G. Lohman, R. Mueller, P. Volk., “GPU Join Processing Revisited”, *Eighth International Workshop on Data Management on New Hardware* (**DaMoN’12**), Scottsdale, Arizona, May 21, 2012.

C12. T. Kaldewey, S. Tata, E. Shekita., “Clydesdale: Structured Data Processing on MapReduce”, *15th International Conference on Extending Database Technology* (**EDBT’12**), Berlin, GERMANY, March 27–30, 2012.

C11. C. Kim, Jatin C., N. Satish, E. Sedlar, A. Nguyen, T. Kaldewey, V. Lee, S. Brandt, P. Dubey, “FAST: Fast Architecture Sensitive Tree Search on Modern CPUs and GPUs”, *2010 ACM SIGMOD/PODS Conference* (**SIGMOD’10**), Indianapolis, IN, June 6–11, 2010. ***Best paper award***

C10. C. Kim, E. Sedlar, J. Chhugani, T. Kaldewey, A. Nguyen, A. Di Blas, V. Lee, N. Satish, P. Dubey, “Sort vs. Hash Revisited: Fast Join Implementation on Modern Multi-Core CPUs”, *35th International Conference on Very Large Databases* (**VLDB’09**), Lyon, FRANCE, August 24–28, 2009

C9. T. Kaldewey, A. Di Blas, J. Hagen, and E. Sedlar, “Parallel Search on Video Cards”, *1st USENIX Workshop on Hot Topics in Parallelism* (**HotPar’09**), Berkeley, California, March 30–31, 2009

C8. T. Kaldewey, A. Di Blas, J. Hagen, E. Sedlar, and S. Brandt, “Memory Matters”, *Work in Progress in the 29th IEEE Real-Time Systems Symposium* (**RTSS’08**), Barcelona, SPAIN, November 30 – December 3, 2008.

C7. Scott Brandt, Carlos Maltzahn, Anna Povzner, Roberto Pineiro, Andrew Shewmaker, and Tim Kaldewey, “An Integrated Model for Performance Management in a Distributed System” *Workshop on Operating Systems Platforms for Embedded Real-Time applications* (**OSPERT’08**), Prague, Czech Republic, July 1 2008

C6. T. Kaldewey, T. M. Wong, R.A. Golding, A. Povzner, and S. Brandt, "Virtualizing Disk Performance", *14th IEEE Real-Time and Embedded Technology and Applications Symposium* (**RTAS'08**), Saint Louis, Missoury, April 22–24, 2008. ***Best paper award***

C5. A. Povzner, T. Kaldewey, S. Brandt, R. Golding, T. Wong, and C. Maltzahn, "Efficient Guaranteed Disk Request Scheduling with Fahrrad", *2008 Eurosys conference* (**Eurosys'08**), Glasgow, UK, March 31 – April, 4 2008

C4. D. Bigelow, S. Iyer, T. Kaldewey, R. Pineiro, A. Povzner, S. Brandt, R. Golding, T. Wong, C. Maltzahn, "End-to-end performance management for scalable distributed storage", *Petascale Data Storage Workshop Supercomputing* (**PDSW'07**), Reno, Nevada, November 11, 2007.

C3. C. Lin, T. Kaldewey, A. Povzner, and S. Brandt, "Diverse Soft Real-Time Processing in an Integrated System", *27th IEEE Real-Time Systems Symposium* (**RTSS'06**), Rio de Janeiro, BRAZIL, December 5–8, 2006.

C2. T. Kaldewey, C. Lin, and S. Brandt, "Firm Real-Time Processing in an Integrated Real-Time System", *Work in Progress in the 12th IEEE Real-Time and Embedded Technology and Applications Symposium* (**RTAS'06**), San Jose, California, April 4–7, 2006.

C1. P. Felber, T. Kaldewey, and S. Weiss, "Proactive Hot Spot Avoidance for Web Server Dependability", *23rd IEEE Symposium on Reliable and Distributed Systems* (**SRDS'04**), Florianpolis, BRAZIL, October 18–20, 2004.

Demonstration of Prototypes

D5. T. Kaldewey, D. K. Tam, D. Wendt, "Accelerating Watson Workloads – Enhancing, scaling, and accelerating human expertise.", *International Conference for High Performance Computing, Networking, Storage and Analysis* (**SC15**), Austin, Texas, November 15–20, 2015.

D4. R. Mueller, T. Kaldewey, G.Lohman, J. McPherson, "WoW: What the World of (data) Warehousing can learn from the World of Warcraft.", *2013 ACM SIGMOD/PODS Conference* (**SIGMOD'13**), New York, New York, June 22–27, 2013.

D3. T. Kaldewey, R. Mueller, G.Lohman, J. McPherson, "GPUs Accelerating SQL – Let your GPU do the heavy lifting in your data warehouse.", *GPU Technology Conference* (**GTC'13**), San Jose, California, March 18–21, 2013.

D2. S. Tata, T. Kaldewey, and A. Balmin, "Clydesdale: Structured Data Processing on Hadoop.", *2012 ACM SIGMOD/PODS Conference* (**SIGMOD'12**), Scottsdale, Arizona, May 20–24, 2012.

D1. T. Kaldewey, R. Mueller, G.Lohman, John McPherson, "GPUs Accelerating SQL – What the World of (data) Warehousing can learn from the World of Warcraft", *Information On Demand* (**IOD'12**), Las Vegas, Nevada, October 21–25, 2012.

Journal Papers

J5. M. Gschwind, T. Kaldewey, D. K. Tam, “Optimizing Efficiency of Deep Learning Through Accelerator Virtualization”, **IBM Journal of Research and Development**, Volume 61, Issue 4/5, 2017.

J4. C. Kim, J. Chhugani, N. Satish, E. Sedlar, A. D. Nguyen, T. Kaldewey, V. W. Lee, S. A. Brandt, and P. Dubey, “Designing fast architecture-sensitive tree search on modern multicore/many-core processors” **ACM Transactions on Database Systems (TODS)** Volume 36, Issue 4, pp. 1–34, December 2011.

J3. A. Di Blas, T. Kaldewey, “Data Monster”, **IEEE spectrum**, Volume 46, Issue 9, pp. 46–51, September 2009.

J2. T. Kaldewey, “Programming Video Cards for Database Applications”, **USENIX ;login**, Volume 34, Issue 4, pp. 21–34, August 2009.

J1. A. Povzner, T. Kaldewey, S. Brandt, R. Golding, T. Wong, and C. Maltzahn, “Efficient Guaranteed Disk Request Scheduling with Fahrrad”, **ACM SIGOPS Operating Systems Review**), Volume 42, Issue 4, pp. 13–25 , May 2008.

Books

B1. T. Kaldewey, A. Di Blas “Large-Scale GPU Search”, **GPU Computing Gems – Jade Edition**, Morgan Kaufmann Publishers, 2011.

Patents

P6. G. Lohman, T. Kaldewey, and P. Volk, “Method for Lock-Free Creation of Partitioned Hash Tables in Parallel”, IBM Research, San Jose, California, filed May 06, 2012, granted Dec 13, 2016 **US Patent 9,519,668**

P5. R.Pei, T. Kaldewey, and S. Raiyani, “Incident command post”, SAP Research, Palo Alto, California, filed Nov 22, 2010, granted Jan 8, 2013 **US Patent 8,352,172**

P4. T. Kaldewey, S. Raiyani, R. Pei, and S. Sobol, “System and method for navigating a facility”, SAP Research, Palo Alto, California, filed Aug 7, 2009, granted Jan. 8, 2013 **US Patent 8,352,176**

P3. T. Kaldewey, S. Raiyani, R. Pei, and S. Sobol, “System and method for navigating a facility”, SAP Research, Palo Alto, California, filed Aug 7, 2009, granted Jan. 1, 2013 **US Patent 8,346,472**

P2. T. Kaldewey, S. Raiyani, R. Pei, and S. Sobol, “System and method for navigating a facility”, SAP Research, Palo Alto, California, filed Mar 14, 2006, granted Sept. 8, 2009 **US Patent 7,587,274**

P1. R.Pei, T. Kaldewey, and S. Raiyani, “Incident command post”, SAP Research, Palo Alto, California, filed Mar 28, 2005, granted Feb 1, 2011 **US Patent 7,881,862**

White Papers

W1. S. Raiyani, R. Pei and T. Kaldewey, “Innovative Architectures for Unified Incident Command and Decision Support”, **White Paper, SAP Research**, Palo Alto, California , October, 2004.

PROFESSIONAL ACTIVITIES

Membership in Professional Societies

ACM since 2008–2014
IEEE since 2005–2014
USENIX since 2009–2014

Program Committee

*8th IEEE International Conference on Networking, Architecture, and Storage (***NAS'13***),*
Xi'An, China, July 17–19, 2013
*18th IEEE Real-Time and Embedded Technology and Applications Symposium (***RTAS'12***),*
Beijing, China, April 16–19, 2012
*6th IEEE International Conference on Networking, Architecture, and Storage (***NAS'11***),*
Dalian, China, July 28–30, 2011

Reviewer

1st IBM **Deep Learning Workshop**,
Yorktown Heights, New York, January 29, 2016.
3rd ACM Symposium on Cloud Computing (**SoCC'12**),
San Jose, California, October 14–17, 2012
GPU Computing Gems – *Emerald Edition*,
Morgan Kaufmann Publishers, 2011
36th International Conference on Very Large Data Bases (**VLDB'10**),
Singapore, September 13–17, 2010
28th IEEE International Conference on Distributed Computing Systems (**ICDCS'08**),
Beijing, China, June 17–20, 2008
5th USENIX Conference on File and Storage Technologies (**FAST'07**),
San Jose, California, February 13–16, 2007
12th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (**RTCSA'06**),
Sydney, Australia, August 16–18, 2006
18th Euromicro Conference on Real-Time Systems (**ECRTS'06**),
Dresden, Germany, July 5–7, 2006

UNIVERSITY SERVICE

Columbia University

2016 Ph.D. committee member for Evangelia Sitaridi, Thesis: *GPU-Acceleration of In-Memory Data Analytics*

Institute Eurecom

2003-2004 Students' Representative

Department of Management Science, University of Technology Darmstadt

2002-2003 Committee Member for Faculty Recruitment

1997-2003 Students' Representative

1998-2000 Apprentice's Chairman