Mehul A. Shah

mashah@gmail.com
(510) 793-4391
12543 Palmtag Drive, Saratoga, CA 95070
http://www.hpl.hp.com/personal/Mehul_Shah/

Interests: Distributed computing, databases – analytics and transactional systems, energy-efficient systems
Key Projects: TelegraphCQ, JouleSort, Sinfonia, HP-KVS, Armonia

Education

University of California, Berkeley
Ph.D., October 2004, EECS Department. Database group
Thesis: *Flux: A Mechanism for Building Robust, Scalable Dataflows*
Advisor: Joseph M. Hellerstein

Massachusetts Institute of Technology
MEng, June 1997, Electrical Engineering and Computer Science
Thesis: *ReferralWeb: A Resource Location System Guided by Personal Relations*
Advisors: David R. Karger, Henry Kautz, and Bart Selman

B.S., June 1996, Computer Science
B.S., June 1996, Physics

Employment History

Hewlett-Packard Laboratories, *October 2004 to present* — Palo Alto, CA
Senior Research Scientist, 2008 to present
Research Scientist, 2004-2008

University of California, **Berkeley**, *September 1997 to October 2004* — Berkeley, CA
Database Group, *Graduate Student Researcher*

IBM Almaden Research Center, *January 1999 to October 1999* — San Jose, CA
DB2/OSF Group**,** *Intern*

AT&T Laboratories – Research, *June 1996 to January 1997* — Murray Hill, NJ
MEng, Thesis Internship

AT&T Bell Laboratories — NJ
Research Intern, *Summer 1995*, Murray Hill, NJ
Research Intern, *Summer 1994*, Holmdel, NJ

Research Contributions

My research spans multiple fields from data management to systems to architecture. I often transfer ideas from one into another and find the most interesting insights and results lie at the intersection of these fields.

Scalable distributed systems: A major theme in my career is of building ever more scalable and robust storage and data management platforms. I am currently a principal investigator for the Armonia project, a memory-centric database which aims to meet the modern demands of workloads like financial trading, telecommunications, social networking, and so on. These applications offer little request locality, need stable low-latency responses, need scalability, and need high availability. To meet these needs, Armonia's key approach is to keep all data in data-structures distributed across the main-memory of a cluster and use Sinfonia to scale these data-structures and keep them consistent and fault-tolerant. Further, we are investigating new interfaces to support predictive and continuous analytics over this data.

Sinfonia is a data-sharing platform that makes it easy to build scalable, distributed, and fault-tolerant infrastructure applications [TOCS09]. As a key contributor to the Sinfonia system, I was intimately involved with developing a group communication system and distributed B-tree on top of Sinfonia. Both are the largest practical implementations of those known to date, scaling well to 100s of nodes. Sinfonia won best paper at SOSP 2007, and has inspired further academic research, e.g. RamCloud.

I am also a founder and key contributor to the HP-KVS project. HP-KVS is an erasure coded, eventually-consistent, self-healing key-value store. The goal of HP-KVS is to provide an object storage service, similar to Amazon's S3, that stores 100s of PBs at lower cost, provides higher availability (> 4 nines), and spans multiple geographies. I implemented the core self-healing, eventually-consistent protocols [DSN10], and developed generic techniques for analyzing the consistency provided by key-value stores [HotDep10, PODC11].

Finally, in the TelegraphCQ system, my doctoral dissertation focused on making parallel CQ dataflows – computations that analyze high-throughput streaming data in real time – highly available and automatically load-balancing.

Energy-efficient database systems: My work in energy-efficiency is focused around characterizing and optimizing the energy use of computer systems as a whole, from storage to memory to compute. It borrows popular workloads from the database community to do so.

I conceived and am the maintainer of the JouleSort benchmark [SIGMOD07], the first holistic energy efficiency benchmark, which has inspired new designs of database servers and influenced other benchmarks. JouleSort is part of the official suite of external sort benchmarks (www.sortbenchmark.org) from the DB community, and has received numerous entries from across the globe. Since its release in 2007, we have seen an 84% per year improvement in energy efficiency of the best sorting machines, a rate that outpaces Moore's law and eclipses the prior 22% per year improvement rate. Moreover, since its release, we have seen other benchmark councils like TPC and SPEC follow suit with energy-efficiency metrics of their own, and increased academic work from various institutions on database server energy efficiency.

Beyond sorting, my research investigates the energy efficiency of more complex database workloads like TPC-H on large scale-up servers, on single node scale-out servers, and across clusters. Contrary to previous studies, we find that on single node scale-out servers, there is little opportunity for software knobs to affect energy efficiency separately from performance; the fastest configuration is also most energy efficient [SIGMOD10]. On large multi-chip scale-up servers [ISLPED09] and across clusters, however, we find software can be used to tune energy use separately from performance by changing physical data layout and leveraging heterogeneity in a cluster.

Leveraging non-volatile RAM technologies: With the availability of practical, fast solid-state drives (SSDs) and the promise of new non-volatile technologies like PC-RAM and Memristor, my work investigates changes in software and hardware stack needed to make most effective use of these devices. I have investigated new storage layouts and algorithms to improve the performance of database query processing over SSDs [SIGMOD09]. I have developed methods for making memory-resident transactional applications durable with little overhead using SSDs. My work also proposes new OS mechanisms needed to build an effective NVRAM – DRAM hybrid memory system [HotOS09]. Finally, I am currently investigating the implications of nanostores, a radical re-design that reduces the memory hierarchy to a single layer of NVRAM coupled with simple cores to form a single scale-out unit. Our initial results show that for modern workloads this simplification of the system stack will provide orders of magnitude improvements in resource efficiency needed to take us into the next computing era.

Other work: Two additional research threads of mine are worth noting. First, I worked on methods for assessing, estimating, and improving the longevity of digitally preserved data [EuroSys06, HotOS07,PP08]. Second, I built ReferralWeb, the first system to automatically generate social networks by mining publicly available data on the web [CACM97, AI97]. The goal of the system was to find experts on particular topics and use one's social network to connect with the experts. Using this tool, I automatically gathered a 1200 person network for my master's thesis – about a decade before the proliferation of online social networking sites.

Awards and Recognition

- 4 HP eAwards
- Best paper SOSP 2007, "Sinfonia: a new paradigm for building scalable distributed systems."
- Siebel Scholars Fellowship, 2003, awarded to U.C. Berkeley EECS grad student with high GPA
- U.C. Microelectronics Fellowship, 1997
- Henry Ford II Scholar Award, 1996, awarded to top performing graduating undergraduate at MIT

Panels and Invited Talks

- High Performance Transaction Systems, "Eventually consistent is eventually not enough," *Oct. 2011,* **To be delivered**.
- U.C. San Diego, Center for Networked Systems review, invited talk, "Cloudy pains," *Feb 2011*.
- DaMoN panel, "How to trust cloud storage," *June 2010*.
- Stanford Infolab workshop panel, topic: Cloud + Transactions, "No SQL is no good," *April 2010*.
- Sonoma State University, CS Colloquim guest lecture, "Benchmarking and designing for energy efficiency," *Oct. 2009.*
- U.C. Berkeley, course CS286 invited guest lecture, "Benchmarking and designing for energy efficiency," *Feb. 2009.*
- SIGMOD New Researcher Symposium, invited talk, "Tips for starting a research career," *June 2008.*

Selected Press

- Slashdot, "Benchmarking power-efficient servers," *August 21, 2007.*
- StorageMojo blog, "Benchmarking energy efficiency," *August 19, 2007*.

Publications

Conference acceptance rates: SIGMOD (< 17%), VLDB (<17 %), SOSP (< 17%), PODC , HotOS (< 15%), DSN (<20%)

Refereed papers

1. [HPTS11] Mehul A. Shah, "Eventual consistency is eventually not enough." position paper, High Performance Transaction Systems, 2011. To be delivered.
2. [PODC11] Wojciech M. Golab, Xiaozhou Li, Mehul A. Shah, "Analyzing consistency properties for fun and profit." *PODC, 2011*.
3. [HotDep10] Eric Anderson, Xiaozhou Li, Mehul A. Shah, Joseph Tucek, and Jay J. Wylie, "What consistency does your key-value store *actually* provide?" *Workshop on Hot Topics in System Dependability, 2010.*
4. [DSN10] Eric Anderson, Xiaozhou Li, Arif Merchant, Mehul A. Shah, Kevin Smathers, Joseph Tucek, Mustafa Uysal, Jay J. Wylie, "Efficient eventual consistency in Pahoehoe, an erasure-coded key-blob archive." *DSN, 2010.*
5. [SIGMOD10] Dimitris Tsirogiannis, Stavros Harizopoulos, Mehul A. Shah, "Analyzing the energy efficiency of a database server." *SIGMOD, 2010*.
6. [TOCS09] Marcos K. Aguilera, Arif Merchant, Mehul A. Shah, Alistair Veitch, Christos Karamanolis, "Sinfonia: A new paradigm for building scalable distributed systems." *Transactions on Computer Systems, Nov. 2009*.
7. [ISLPED09] Justin Meza, Mehul A. Shah, Parthasarathy Ranganathan, Mike Fitzner, and Judson Veazey. "Tracking the power in an enterprise decision support system." *International Symposium on Low Power Electronics and Design, August 2009*.
8. [SIGMOD09] Dimitris Tsirogiannis, Stavros Harizopoulos, Mehul A. Shah, Janet L. Wiener, Goetz Graefe, "Query processing techniques for solid state drives." *SIGMOD,2009*.
9. [HotOS09] Jeffrey C. Mogul, Eduardo Argollo, Mehul A. Shah, Paolo Faraboschi: "Operating System Support for NVM+DRAM Hybrid Main Memory." *Workshop on Hot Topics in Operating Systems, 2009*.
10. [CIDR09] Stavros Harizopoulos, Mehul A. Shah, Justin Meza, Parthasarathy Ranganathan, "Energy efficiency: The new holy grail of data management systems research." *Conference on Innovative Data Systems Research, 2009.*
11. [VLDB08] Marcos K. Aguilera, Wojciech M. Golab, Mehul A. Shah, "A practical scalable distributed B-tree." *PVLDB, 2008.*
12. [HotDep08] Amitanand S. Aiyer, Eric Anderson, Xiaozhou Li, Mehul A. Shah, Jay J. Wylie, "Consistability: Describing usually consistent systems." *Workshop on Hot Topics in System Dependability, 2008*.
13. [DaMoN08] Mehul A. Shah, Stavros Harizopoulos, Janet L. Wiener, Goetz Graefe, "Fast scans and joins using flash drives." *Workshop on Data Management on New Hardware, 2008.*
14. [SOSP07] Marcos K. Aguilera, Arif Merchant, Mehul A. Shah, Alistair C. Veitch, Christos T. Karamanolis, "Sinfonia: a new paradigm for building scalable distributed systems." *SOSP, 2007*. **Best paper. (87 citations on Google scholar)**
15. [SIGMOD07] Suzanne Rivoire, Mehul A. Shah, Parthasarathy Ranganathan, Christos Kozyrakis, "JouleSort: a balanced energy-efficiency benchmark." *SIGMOD, 2007.* **(101 citations on Google scholar)**
16. [HotOS07] Mehul A. Shah, Mary Baker, Jeffrey C. Mogul, Ram Swaminathan, "Auditing to Keep Online Storage Services Honest." *Workshop on Hot Topcs in Operating Systems, 2007*.
17. [NSDI06] Patrick Reynolds, Charles E. Killian, Janet L. Wiener, Jeffrey C. Mogul, Mehul A. Shah, Amin Vahdat,

Pip: Detecting the Unexpected in Distributed Systems." *NSDI, 2006*.

18. [EuroSys06] Mary Baker, Mehul A. Shah, David S. H. Rosenthal, Mema Roussopoulos, Petros Maniatis, Thomas J. Giuli, Prashanth P. Bungale, "A fresh look at the reliability of long-term digital storage." *EuroSys, 2006*.
19. [SIGMOD04] Mehul A. Shah, Joseph M. Hellerstein, Eric A. Brewer, "Highly available, fault-tolerant, parallel dataflows." *SIGMOD, 2004*. **(89 citations Google scholar)**
20. [ICDE03] Mehul A. Shah, Joseph M. Hellerstein, Sirish Chandrasekaran, Michael J. Franklin, "Flux: An adaptive partitioning operator for continuous query systems." *ICDE 2003*. **(173 citations Google scholar)**
21. [SIGMOD02] Samuel Madden, Mehul A. Shah, Joseph M. Hellerstein, Vijayshankar Raman, "Continuously adaptive continuous queries over streams." *SIGMOD, 2002.* **(575 citations Google scholar)**
22. [UIDIS99] Mehul A. Shah, Marcel Kornacker, Joseph M. Hellerstein: Amdb: A Visual Access Method Development Tool. *Workshop on User Interfaces to Data Intensive Systems, 1999.*

Book chapters

1. [Book09] Magdalena Balazinska, Jeong-Hyon Hwang, Mehul A. Shah, "Fault-tolerance and high availability in data stream management systems." *Encyclopedia of Database Systems, 2009*: 1109-1115.

Magazine articles and tech reports

1. [Computer07] Suzanne Rivoire, Mehul A. Shah, Parthasarathy Ranganathan, Christos Kozyrakis, Justin Meza, "Models and metrics to enable energy-efficiency optimizations." *IEEE Computer, 2007*, 40(12): 39-48.
2. [Pervasive06] Ajay Gupta, Parthasarathy Ranganathan, Prashant Sarin, Mehul A. Shah, "IT infrastructure in emerging markets: Arguing for an end-to-end perspective." *IEEE Pervasive Computing, 2006*, 5(2): 24-31.
3. [AI97] Henry A. Kautz, Bart Selman, Mehul A. Shah: The Hidden Web. *AI Magazine, 1997*, 18(2): 27-36.
4. [CACM97] Henry A. Kautz, Bart Selman, Mehul A. Shah, "Referral Web: Combining social networks and collaborative filtering." *CACM, 1997*, 40(3): 63-65. **(684 citations Google scholar)**
5. [PP08] Mehul A. Shah, Ram Swaminathan, Mary Baker, "Privacy-preserving audit and extraction of digital contents." HP Labs technical report, HPL-2008-32R1, 2008.

HP TechCon papers and posters

1. Eric Anderson, Xiaozhou (Steve) Li, Mehul A. Shah, Kevin Smathers, Joseph Tucek, Alistair Veitch, Jay J. Wylie, Bob Souza, "Stout: A highly-available key-value service for the cloud." April 2011. Paper.
2. Jichuan Chang, Parthasarathy Ranganathan, Gilberto Ribeiro, David Roberts, Mehul A. Shah, John Sontag, Jieming Zhu, "Memristor-based datacentric data centers." May 2010. Paper. **Selected as one of HP's most promising technologies**.
3. Eric Anderson, Xiaozhou Li, Mehul A. Shah, Kevin Smathers, Joseph Tucek, Jay J. Wylie, Amit Sharma, "Pahoehoe: Highly Available, key-blob storage for the cloud." May 2010. Poster.
4. Eric Anderson, Xiaozhou Li, Mehul A. Shah, Joseph Tucek, Jay J. Wylie, Michael Callahan, James Pownell, Amit Sharma, Mark Watkins, "A key-value store for services in the cloud." May 2009. Honorable mention invitation.

Patents granted

1. *7,647,454*. Marcos K. Aguilera, Christos Karamanolis, Arif Merchant, Mehul A. Shah, Alistair Veitch, "Transactional shared memory system and method of control."
2. *7,609,703*. Mehul A. Shah, Marcos K. Aguilera, Christos Karamanolis, Arif Merchant, Alistair Veitch, "Group communication system and method."

Patents pending

1. 201001570, with Joseph Tucek, Eric Anderson, "Systems And Methods For Fine Granularity Memory Sparing," April, 2011.
2. 201001937, with Xiaozhou Li, "Directed Graphs Pertaining to Read/Write Operations," Dec. 2010.
3. 200904328, with Jichuan Chang, Parthasarathy Ranganathan, "Accessing A Local Storage Device Using An Auxiliary Processor," August, 2010.
4. 201000086**,** with Jichuan Chang, Parthasarathy Ranganathah, David Roberts, John Sontag, **"**Apparatus Having A

Flattened-Level Data Storage Hierarchy And Methods For Its Use," March, 2010.

5. 200903090**,** with Eric Anderson, Xiaozhou Li, Jay Wylie, **"**Recovery Procedure For A Data Storage System," January, 2010.
6. 200903341, with Eric Anderson, Xiaozhou Li, Jay Wylie, "Scrubbing Procedure For A Data Storage System," January, 2010.
7. 200902335, with Jeff Mogul, Eduardo Argollo, Paolo Faraboschi, "Main Memory With Non-volatile Memory And DRAM," September, 2009.
8. 200900057, with Dimitris Tsirogiannis, Janet Wiener, Stavros Harizopoulos, Goetz Graefe, "Fetching Optimization In Multi-way Pipelined Database Joins," May, 2009.
9. 200802048, with Janet Wiener, Stavros Harizopoulos, Goetz Graefe, "Database Join Optimized For Flash Storage," February, 2009.
10. 200802974, with Amit Aiyer, Eric Anderson, Xiaozhou Li, Jay Wylie, "Methods Of Measuring Consistability Of A Distributed Storage System," January, 2009.
11. 200801128, with Rob Schreiber, Alina Ene, Nikola Milosavljevic, Robert Tarjan, "Computer-implemented Method For Obtaining A Minimum Biclique Cover In A Bipartite Dataset," January, 2009.
12. 200700296, with Rob Schreiber, Robert Tarjan, William Horne, "Computer Implemented Method For Role Discovery In Access Control Systems," August, 2007.
13. 200603747, with Marcos K. Aguilera, Wojciech Golab. "Providing A Distributed Balanced Tree Across Plural Servers," April, 2007.

Miscellaneous

Committees

- Sort benchmark committee, review and audit entries annually, *2007 to present.*
- SIGMOD PC, *2010, 2009*
- VLDB PC, *2011 (Industrial), 2007*
- ICDE PC, *2007*
- HotPower PC, *2009*
- Reviewer for: FAST, SIGMETRICS, PODC, OSDI, SOSP, Transactions on Storage (*2010*), Transactions on Database Systems (*2011, 2004*)
- Co-chair, HP Labs ECSR workshop, *December 2007*

Mentoring

- Summer students: James Anderson (UCSD, 2006), Suzanne Rivoire (Stanford, 2006-7), Justin Meza (CMU, 2007-9), Mohit Saxena (U. Wisconsin, 2010)
- PhD thesis committee: Risi Thonangi, Duke University. Advisor: Jun Yang

Software released: TelegraphCQ (http://telegraph.cs.berkeley.edu/), Amdb (http://gist.cs.berkeley.edu/)

Hobbies: Running and photography