Minglong Shao
Computer Science Department, 5000 Forbes Ave, Pittsburgh, PA 15213
Office: (412) 268-7669 Mobile: (412) 608-4965
shaoml@cs.cmu.edu http://www.cs.cmu.edu/~shaoml

OBJECTIVE

Research position in industry or academia

RESEARCH INTERESTS

Computer system, especially database systems and their interaction with underlying processor and memory hierarchy, as well as storage devices

EDUCATION

Ph.D. of computer science, Carnegie Mellon University, U.S.A.	2001-2007
Master student of computer science, Tsinghua University, China	2000-2001
Bachelor of computer science, Tsinghua University, China	1995-2000

HONORS & AWARDS

Research Fellowship at Carnegie Mellon University	2001-2007
USENIX conference scholarship (FAST '04)	2004
"Outstanding Graduating Student", Tsinghua University	2000
Various prestigious scholarships in Tsinghua University, annually	1995-2001

EMPLOYMENT

Summer Intern at IBM Almaden Research Center, San Jose	2005
Summer Intern at Intel Research Pittsburgh, Pittsburgh	2002

PUBLICATIONS

- Minglong Shao, Stratos Papadomanolakis, Steven Schlosser, Jiri Schindler, Anastassia Ailamaki, Gregory Ganger. *MultiMap: Preserving disk locality for multidimensional datasets*. To appear at International Conference of Data Engineering 2007
- Quanzhong Li, Minglong Shao, Volker Markl, Kevin Beyer, Latha Colby, Guy Lohman. *Adaptively reordering joins during query execution*. To appear at International Conference of Data Engineering 2007
- Steven Schlosser, Jiri Schindler, Stratos Papadomanolakis, Minglong Shao, Anastassia Ailamaki, Christos Faloutsos, Gregory Ganger. *On multidimensional data and modern disks*. USENIX Conference on File and Storage Technologies 2005.
- Minglong Shao, Anastassia Ailamaki, Babak Falsafi. *DBmbench: Fast and accurate database workload representation on modern microarchitecture*. Annual International Conference of Computer Science and Software Engineering 2005.
- Minglong Shao, Jiri Schindler, Steven Schlosser, Anastassia Ailamaki, Gregory Ganger. *Clotho: Decoupling memory page layout from storage organization*. International Conference on Very Large Data Bases 2004.
- Jiri Schindler, Steven Schlosser, Minglong Shao, Anastassia Ailamaki, Gregory Ganger. *Atropos: A disk array volume manager for orchestrated use of disks*. USENIX Conference on File and Storage Technologies 2004.
- Changhao Jiang, Minglong Shao, Yiheng Li, Peng Jia, *Accelerating clustering methods through fractal based analysis*. Fractal workshop (in conjunction with SIGKDD) 2002.

RESEARCH EXPERIENCE

- **Preserve disk locality for multidimensional datasets (2004 – 2005)**
  The goal of this project is to solve the fundamental problem of preserving spatial locality for multidimensional datasets. We first augment the existing disk linear interface with a new adjacency model that provides a new multidimensional view for disks. This model uses existing technology and functions readily available inside disk firmware to identify non-contiguous logical blocks that can be accessed with minimal positioning cost. Then we propose a new mapping model called MultiMap based on the adjacency model to map multidimensional datasets to the linear address space of storage systems.

MultiMap provides full streaming bandwidth for one (primary) dimension and maximally efficient non-sequential access (i.e., minimal seek and no rotational latency) for the other dimensions by preserving disk locality for multidimensional datasets. Experimental evaluation of a prototype implementation demonstrates MultiMap's superior performance for range and beam queries. Project home page: http://www.pdl.cmu.edu/Database/multimap.html

- **Dynamic reordering of nested loop joins (Summer intern 2005 at IBM)**
  Dynamic query re-optimization has been proposed as an effective solution to detect and correct query optimizer errors due to incorrect or unavailable statistics or simplified cost metrics promptly during the query execution. Dynamic reordering of nested loop joins is a light-weight runtime re-optimization technique targeting at rearranging the order of join tables of the left-deep nested join pipeline according to the latest observed data characteristics. It leverages the pipelined join process to minimize the bookkeeping overhead and to avoid the re-computation of processed tuples. By paying little cost (space-wise and time-wise), this approach provides us quite a chance to re-optimize query plans. It achieves its goal in two ways: 1. detects and corrects the bad outermost table; 2. adapts to the detailed, changing data distributions of join tables. Experiment results on our prototype show that dynamic query re-optimization is a simple, yet effective solution.
- **Fates database storage manager (Summer 2003 – 2004)**
  The goal of the Fates architecture is to offer efficient execution at all levels of memory hierarchy and optimize data layout to improve performance, by exploiting the unique characteristics available at each level. This is primarily done by decoupling the in-memory data layout from the storage organization. Where traditional database systems are forced to fetch and store unnecessary data as an artifact of a chosen data layout, the Fates database system can request, retrieve, and store just the needed data, catering to the needs of a specific query. This conserves storage device bandwidth, memory capacity, and avoids cache pollution, thereby significantly improving both query execution time and tightening the interaction between the database software and the underlying storage and memory subsystem. The prototype includes three parts: Clotho, Lachesis, and Atropos. Project home page: http://www.pdl.cmu.edu/Database/fates.html.
- **Database Microbenchmarking (Fall 2002-Fall 2003)**
  Designed and evaluated DBmbench, a database microbenchmark suite, to mimic the performance behavior of full-scale DSS and OLTP workloads at the computer micro-architectural level where conventional database benchmarks are inapplicable due to the longer execution times which are usually orders of magnitude larger on simulators than on real machines.
- **Course Project of Multimedia Database and Data Mining (Spring 2002)**
  The correlation integral plot of a dataset has implicit information to estimate the natural number of clusters in the dataset. It also might be used to determine the critical sampling size of a subset with preserved data distribution. The motivation of the course project is to utilize the information provided by the correlation integral plot to improve performance of existing clustering methods (BIRCH, CURE and OPTICS).

TEACHING EXPERIENCE

Teaching assistant of " Introduction to computer systems", Carnegie Mellon University	Fall 2004
Teaching assistant of " Database applications", Carnegie Mellon University	Spring 2003

REFERENCE

Available upon request