Yuzhe (Richard) Tang

Addr: 329142, Atlanta, GA 30332, USA
Phone: +1-678-793-2706
Email: yztang@gatech.edu
Web: http://www.cc.gatech.edu/~ytang36/

Objective

Seeking a research-oriented position in academia or industry labs.

Research Interests

My research interests are on the secure distributed systems for big data management, with a current focus on *indexing key-value stores, outsourcing key-value stores with authentication, privacy-aware search by optimizing multi-party computations.*

Education

Georgia Institute of Technology Aug. 2009 – May 2014 (expected)
Ph.D. in Computer Science Adviser: **Ling Liu**
Thesis: Scalable and Secure Cloud Services in Big Data Systems

Fudan University (in China) Sep. 2006 – Jun. 2009
M.Sc. in Computer Science Adviser: Shuigeng Zhou

Fudan University (in China) Sep. 2002 – Jun. 2006
B.Sc. in Computer Science

Publications

- *First-authored papers*

1. **Yuzhe Tang**, Ling Liu, Arun Iyengar. ϵ-PPI: Searching Information Networks with Quantitative Privacy Guarantees, in **ICDCS 2014**, (Acceptance ratio $< 13\%$).
2. **Yuzhe Tang**, Ting Wang, Xin Hu, Reiner Sailer, Peter Pietzuch, Ling Liu. Outsourcing Key-Value Stores with Verifiable Data Freshness, to appear in **ICDE 2014** (System Demo Track).
3. **Yuzhe Tang**, Bugra Gedik. Auto-pipelining in Data Stream Processing, in **TPDS 2013**, Preprint ISSN: 1045-9219, DOI: 10.1109/TPDS.2012.333.
4. **Yuzhe Tang**, Ting Wang, Ling Liu, Shicong Meng, Balaji Palanisamy. Privacy Preserving Indexing for eHealth Information Networks, in **CIKM 2011**, pages 905 – 914, (Acceptance ratio $< 14\%$).
5. **Yuzhe Tang**, Jianliang Xu, Shuigeng Zhou, Wang-Chien Lee, Dingxiong Deng, Yue Wang. A Lightweight Multi-dimensional Index for Complex Queries over DHTs, in **TPDS 2011**, Vol 22, Issue 12, pages 2046 – 2054.
6. **Yuzhe Tang**, Shuigeng Zhou, Jianliang Xu. LIGHT: A Query-Efficient yet Low-Maintenance Indexing Scheme over DHTs, in **TKDE 2010**, Vol 22, No. 1, pages 59 – 75.
7. **Yuzhe Tang**, Jianliang Xu, Shuigeng Zhou, Wang-Chien Lee. m-LIGHT: Indexing Multi-Dimensional Data over DHTs, in Proceedings of **ICDCS 2009**, pages 191–198, (Acceptance ratio $< 16\%$).
8. **Yuzhe Tang**, Shuigeng Zhou. LHT: A Low-Maintenance Indexing Scheme over DHTs, in Proceedings of **ICDCS 2008**, pages 141–151, (Acceptance ratio $< 16\%$).

- *Collaboration papers*

9. Wei Tan, Sandeep Tata, **Yuzhe Tang**, Liana Fong, Diff-Index: Differentiated Index in Distributed Log-Structured Data Stores, to appear in **EDBT 2014**.
10. Balaji Palanisamy, Ling Liu, Kisung Lee, Shicong Meng and **Yuzhe Tang**, Anonymizing Continuous Queries with Delay-tolerant Mix-zones on Road Networks, In International Journal of Distributed Parallel Databases 2013.

11. Qi Zhang, Ling Liu, Yi Ren, Kisung Lee, **Yuzhe Tang**, Xu Zhao, Yang Zhou. Residency Aware Inter-VM Communication in Virtualized Cloud: Performance Measurement and Analysis, In IEEE Cloud 2013.
12. Kisung Lee, Ling Liu, **Yuzhe Tang**, Qi Zhang, Yang Zhou. Efficient and Customizable Data Partitioning Framework for Distributed Big RDF Data Processing in the Cloud, In IEEE Cloud 2013.
13. Shicong Meng, Arun Iyengar, Isabelle Rouvellou, Ling Liu, Kisung Lee, Balaji Palanisamy, **Yuzhe Tang**. Reliable State Monitoring in Cloud Data Centers, in IEEE Cloud 2012, pages 951 – 958, (**Best Paper Award**).
14. Balaji Palanisamy, Ling Liu, Kisung Lee, Aameek Singh and **Yuzhe Tang**. Attack Resilient Mix-zones for Mobile Objects on Road Networks, in IEEE MSN 2012.

- *Papers under submission*

15. Yuzhe Tang, Ting Wang, Xin Hu, Jiyong Jang, Peter Pietzuch, Ling Liu. Authentication of Freshness for Outsourced Multi-Version Key-Value Stores, in submission.
16. Yuzhe Tang, Arun Iyengar, Wei Tan, Ling Liu, Liana Fong. Write-Optimized Indexing for Log-Structured Key-Value Stores, in submission.
17. Yuzhe Tang, Ling Liu. Multi-Keyword Privacy-Preserving Search in Personal Server Networks, in submission.
18. Yuzhe Tang, Junichi Tatemura, Ling Liu, Hakan Hacigumus. KTV-Tree: Real-time Top-k Analytics by Dynamic View Materialization in Cloud. Tech Report 2010.

Patent

Bugra Gedik, Scott A. Schneider, **Yuzhe Tang**, Kun-lung Wu. Adaptive auto-pipelining in Stream Processing Applications. Applied patent, Disclosure YOR820110835, 2011.

Research Projects

Part 1: Performance Optimization of Scalable Cloud Serving Systems

- **Indexing write-optimized key-value stores [9,16]**

  Key-value stores are emerging big data storage and serving systems which attract increasing amount of attentions. Many write-optimized key-value stores, such as HBase and Cassandra, have been successfully used to manage write-intensive workloads for Web 2.0 applications. While most existing key-value stores provide key-based access methods (e.g. Put/Get), the value-based access methods are rarely supported due to the challenges in indexing big data at scale. In this project, I addressed the indexing of key-value stores under write-intensive workloads, and proposed HIndex, a generic index framework adaptable to various key-value stores. HIndex achieves instant index updates in real time, yet with very lightweight maintenance overhead. My approach is to use a performance optimization technique that is designed aware of the unique characteristic of write-optimized key-value stores. Concretely, I discovered that the inefficiency of using conventional indexing approaches on write-optimized workloads comes from an expensive indexing operation, named index repair. My performance optimization technique [16] defers the index repair operations to a later time, either during an online period when the system is serving the other operations, or in an offline stage when the system is under low workload. The online design of HIndex aims at making index repair as lightweight as possible in order to avoid intruding the performance of concurrent online operations, while the offline design optimizes the throughput of batch processing by a tight coupling of the repair operations with a store reorganization operation called compaction. HIndex is an adaptive indexing framework that adapts the decision making of deferred index repairs to the current system workload. The proposed HIndex technique has been integrated into the IBM BigInsights product [9].

- **Scaling up stream processing with multi-cores by automatic pipelining [3]**

  Stream processing, critical to many big data applications such as real-time analytics, is known to be computationally intensive. To fully utilize the multi-core resources, it is desirable to use the

system optimization techniques. In this project, I studied the problem of automatic performance optimization of streaming applications by using pipelining parallelism [3]. Through modeling the system CPU bottleneck, I formulated that the design goal is to prevent the single-core bottleneck from happening, in which one or few saturated cores may block the whole system. I proposed a program-to-thread mapping scheme that can dynamically adjust itself based on the changing workloads and automatically avoids the single-core bottleneck. I implemented the proposed scheme in the IBM System S which is an industrial strength big stream processing system. The implementation is based on a lightweight CPU profiler and optimizer under an adaptive control framework.

- LIGHT**: Indexing DHT networks with low maintenance overhead [5,6,7,8]**
  This project aims at providing a generic index on top of distributed hash tables (or DHT), which is a representative peer-to-peer network. Due to the high dynamism in peer-to-peer networks where peers can freely join and leave the network, it calls for a lightweight index maintenance scheme. I proposed LIGHT, an optimized indexing framework on top of generic DHT networks. The design of LIGHT challenges two seemingly conflicting goals, that is, maximizing query efficiency while minimizing the maintenance overhead. To meet the challenge, I proposed a novel index-to-peer mapping scheme that intelligently minimizes inter-peer communications for both index maintenance and query processing. I applied this idea to different indexing and search scenarios with the devised algorithms, including value-based range query [6,8], multi-dimensional queries [7] and k-NN queries [5].

Part 2: Security and Privacy in Multi-domain Cloud Systems

- **Authenticated Put/Get in outsourced key-value stores [2,15]**
  This project addresses the data authenticity of an outsourced key-value storage in the cloud. In an outsourced database, the data owner publishes the local data updates to the cloud which is responsible for serving the big data to a large customer base. The outsourced data and updates are signed to protect the data authenticity. In this work, I focused on the multi-version key-value stores and addressed the authentication of version freshness. That is, it is guaranteed that a latest version of an object is returned from the cloud storage. I proposed an authentication framework [15] to sign the data updates in a streaming fashion and verify the data freshness. To optimize the performance of data signing under an intensive data stream, I further proposed INCBM TREE [2], an authentication structure based on a hierarchy of Bloom Filters digested by a Merkle tree. For efficient verification, the INCBM TREE can effectively summarize the data update stream yet it requires a very small memory footprint which lends itself to the owner of limited resources.
- **Privacy preserving index and search in multi-domain clouds [4,1,17]**
  In the age of cloud computing, losing data control continues to be a major concern and the recent move toward giving data control back to cloud users has given birth to a variety of multi-domain cloud systems for different applications such as distributed social networking, peer-to-peer file sharing and electronic Healthcare It is crucial to support privacy preserving index (or PPI) in the multi-domain cloud for the sake of effective information exchange and sharing between domains. In this project, I proposed a number of PPI frameworks addressing different system designs. First, for efficient secure index construction, I proposed ssPPI [4] based on the novel use of secret sharing in a parallel and distributed computing framework. Second, for effective privacy preservation, I proposed ϵ-PPI [1] that differentiates the protection on different indexed terms with quantitatively controllable privacy. Third, to address the privacy protection under multi-keyword document search, I proposed MPPI [17]. The key challenge in realizing the proposed PPI frameworks comes from the needs for secure index construction in a mutually untrusted network. While the norm to use multi-party computations (or MPC) for secure computation is very expensive when it comes to big data in large network, I proposed several MPC optimization techniques to realize the massive computation in the PPI construction on tens or hundreds of cloud domains.

Work Experiences

Georgia Tech, Atlanta, GA USA Aug/2009–present

Lead Research Assistant
This NSF-funded project, entitled "Privacy Preserving Index on Access-controlled Documents", aims at developing a privacy preserving information networks, where access-controlled documents distributed cross multiple mutually-untrusted domains can be efficiently searched without sacrificing privacy. The result of this on-going project is research papers published in CIKM'11 and ICDCS'14.

- Proposed a secret-sharing based mechanism to securely compute the privacy preserving index (i.e. PPI).
- Proposed various computation models for different privacy concerns, including multi-keyword search and differentiated sensitivity.
- Implemented the proposed computation models based on a multi-party computation platform, FairplayMP.

IBM Research T.J. Watson, Yorktown Heights, NY USA — May/2013–Aug/2013
Research Intern
The project, entitled "Authenticating Big Data Streams", is to authenticate data stream in an outsourced-database scenario. One result of this on-going project is a system demo paper published in ICDE'14.

- Proposed a stream signing approach based on Bloom filters and Merkle tree.
- Implemented a prototype system for outsourcing stream based on HBase and Netty.

IBM Research T.J. Watson, Hawthorne, NY USA — May/2012–Aug/2012
Lead Research Intern
The project, entitled "Index Support on NoSQL", is to enrich NoSQL store (e.g. HBase) with secondary index support. One result of this on-going project is a software integrated in the IBM's BigInsights product and a paper published in EDBT'14.

- Proposed a real-time indexing framework that lends itself to the system of key-value stores. The idea is to take advantage of the fast-write-slow-read characteristic of the HBase alike key-value stores.
- Implemented a complete and functioning prototype system based on HBase's CoProcessor Interface.
- Conducted experiments with a Cloud benchmark tool, YCSB and in a Cloud platform, Emulab. The experiments are fully automated.

IBM Research T.J. Watson, Hawthorne, NY USA — May/2011–Sep/2011
Research Intern
The project, entitled "System Optimization for Stream Processing", is to optimize the performance of data stream processing dynamically on a multi-core system. The result of the project is a research paper published in TPDS'13.

- Identified the streaming system bottleneck to be a CPU utilization problem.
- Proposed a dynamic optimization technique by automatic pipelining the data stream to avoid the bottleneck scenario.
- Implemented the proposed algorithm with full integration into the IBM's streaming platform, System S. The implementation requires Linux system programming.

NEC Labs America, Cupertino, CA USA — May/2010–Aug/2010
Research Intern
The project, entitled "Top-k Aggregation in Cloud", is to support interactive query processing for top-k aggregations in the key-value store in cloud.

- Proposed an adaptive algorithm to maintain materialized views of top-k aggregations.

Fudan University, Shanghai China Sep/2006–Sep/2009
Lead Research Assistant
The project, entitled "P2P Data Management", aims at scaling out database over the distributed hash tables (i.e. DHT) in P2P networks. The results of the project are research papers published in TPDS'11, TKDE'10, ICDCS'09 and ICDCS'08.

- Proposed an original mapping mechanism for low-maintenance indexing over DHT without sacrificing query performance.
- Applied the proposed mapping mechanism to various data models including multi-dimensional data and range and k-NN queries, and proposed algorithms for query processing.

Microsoft Research Asia, Beijing China Feb/2009–Jun/2009
Research Intern
The project, entitled "Log Analysis in Data Centers", aims at automatic performance diagnosis and debugging of distributed systems in a large-scaled data center.

- Analyzed the system log and modeled the distributed system behavior by the acyclic function-call graphs.
- Designed various analysis operators (e.g. behavior-diff and clustering) by applying graph mining techniques (e.g., graph edit distances and metric space clustering).

Teaching Experience

CS3310: Operating system design Fall 2011
Teaching Assistant **Georgia Tech**
Responsible for course project design and instruction on topics of Linux kernel programming and system call tracing. Delivered tutorial and in-class demo to familiarize students with projects. Held follow-up sessions to help solve programming problems. Graded assignments and project with assessment of student learning and feedback to instructor.

CS4420: Database Implementation Fall 2013
Teaching Assistant **Georgia Tech**
Responsible for grading and assignment instruction. Delivered tutorial and in-class demo to familiarize students with Cloud platform including Hadoop and Emulab. Held follow-up sessions to give advises on project designs and implementations. Graded assignments and exams.

CS6675: Advanced Internet Computing and Application Development Fall 2012
Teaching Assistant **Georgia Tech**
Responsible for grading weekly reading reports and final technique review.

VB Programming Language Spring 2008
Teaching Assistant **Fudan University**
Responsible for lab instruction. Gave lectures on VB programming, held class in Lab, and graded experiment reports.

Courses (at Georgia Tech)

Applications : Software Analysis and Testing, Computer System Security, Internet Computing and Application Development, Databases

Systems : Advanced Operating Systems, High Performance Computer Architecture

Theory : Randomized Algorithms

Business(Minor) : Principle Management for Engineers, Financial & Managerial Accounting I, Legal Issues-Technology Transfer

Selected Course Projects

- Intercepted system call using PTrace to enforce additional access controls in a FTP server.
- Implemented using secret sharing to securely store mispelling patterns, for tolerating password case mispelling.
- Administrated Hadoop over virtual machines for performance evaluation.
- Implemented Linux kernel module to capture read/write() system call.
- Implemented pointer analysis.

Skills

System: HBase/Hadoop/YCSB/Cassandra, MySQL/JDBC

Tool: ∗nix/POSIX , Emulab, Netty/ProtoBuf, LaTeX

Programming: Java/Ant, C/C++, Bash/Awk/Expect, Perl/Python

Professional Services

Reviewer: TWeb'14, ICDCS'14, TKDE'13, WWW'13, Middleware'12, DEBS'12, JPDC'11, ICDCS'10

Honors and Awards

- Chinese Government Award for Outstanding Self-financed Students Abroad, 2012
- Best paper award, 5th International Conference on Cloud Computing, 2012 (co-recipient).
- Outstanding Master Thesis of Shanghai, Shanghai Government, 2010
- Tung's Oriental Scholarship, Tung's Oriental, 2008
- HP Distinguished Chinese Student Scholarship, Hewlett-Packard, 2008
- ICDCS Student Travel Grant, TCDP (IEEE Computer Society), 2008
- Graduate Student Fellowship of Fudan University, 2007-2008 (2 times)
- Outstanding Graduated Student of Fudan University, 2006
- Excellence Award, Tencent Innovation Contest, 2006
- The People's Scholarship of Fudan University, 2002–2006 (4 times)
- Chinese Physics Olympiads, First Prize in Hunan Province, 2001

References

Prof. **Ling Liu**, Professor, Adviser
College of Computing
Georgia Institute of Technology
Email: lingliu@cc.gatech.edu
Phone: +1-404-385-1139

Prof. **Calton Pu**, Professor and John P. Imlay, Jr. Chair in Software
College of Computing
Georgia Institute of Technology
Email: calton.pu@cc.gatech.edu
Phone: +1-404-385-1106

Prof. **Mustaque Ahamad**, Professor
College of Computing
Georgia Institute of Technology
Email: mustaq@cc.gatech.edu
Phone: +1-404-894-2593

Prof. **Edward Omiecinskia**, Associate Professor
College of Computing
Georgia Institute of Technology
Email: edwardo@cc.gatech.edu