

YOGESH L. SIMMHAN

150 S. Woodlawn Ave., LH 316
Bloomington IN 47405

ysimmhan@cs.indiana.edu
www.simmhan.com

+1 (812) 337-1347 (M)
+1 (812) 855-9145 (W)

Research Interests

My research interest lies in managing distributed data and metadata for large scale, data driven applications based on Grid and web-services paradigms. I am particularly interested in issues related to provenance tracking, discovery, quality evaluation, and long term preservation of data generated from scientific workflows. The overarching goal of my research is to build scalable information systems that effectively manage data end-to-end in distributed, collaborative environments to support knowledge discovery.

Education

- **Ph.D. Computer Science** (*August 2007*)
Indiana University, Bloomington, IN, USA
 - Thesis title: *Meaningful Quality Metrics for Establishing User Trust in Derived Data*
 - Minor: Software Engineering
 - Advisers: Prof. Beth Plale and Prof. Dennis Gannon
- **M.S. Computer Science** (*May 2002*)
Indiana University, Bloomington, IN, USA
 - GPA: 3.94/4.0
- **B.E. Computer Science and Engineering** (*May 2000*)
University of Madras, S.R.M. Engineering College, Chennai, India
 - Aggregate score: 82% (GPA Equivalent: 3.75/4.0)

Professional Appointments

- **Research Assistant** (*January 2001 - Present*)
Computer Science Department, Indiana University, Bloomington, IN
 - Working with Prof. Beth Plale and Prof. Dennis Gannon on distributed computing, web and grid services middleware, science portals, data provenance, and distributed data and metadata management.
- **Research Co-op** (*May - August 2004*)
IBM Almaden Research Center, San Jose, CA
 - Worked with Inderpal Narang and Vitthal Gogate in the *Data Grid group*. Performed a study of techniques to monitor and improve the quality of service of DB2 Data Replication for on-demand information systems.

- **Research Co-op** (*May - August 2003*)
IBM Almaden Research Center, San Jose, CA
 - Worked with Berthold Reinwald and Jussi Myllymaki in the *Advanced Data Services group*. Worked on XML tools to enhance DB2's support for business process integration, workflow monitoring, and information dissemination.

Teaching Appointments

- **Co-Instructor, Distributed Systems** (*Spring 2007*)
Computer Science Department, Indiana University, Bloomington, IN
 - Teaching graduate level Distributed Systems course with Prof. Beth Plale for a class of 35 students.
- **Co-Instructor, Data Structures** (*Spring 2006*)
Computer Science Department, Indiana University, Bloomington, IN
 - Taught undergraduate level Data Structures course along with Prof. Dennis Gannon for a class of 30 students.
- **Associate Instructor, Programming Concepts using Java** (*Fall 2000*)
Computer Science Department, Indiana University, Bloomington, IN
 - Assisted faculty and lead lab sessions for 8 week undergraduate level course having 40 students.
- **Associate Instructor, Survey of Computers and Computing** (*Fall 2000*)
Computer Science Department, Indiana University, Bloomington, IN
 - Assisted faculty and lead lab sessions for 8 week undergraduate level course having 40 students.
- **Instructor, Java Programming** (*July 1999*)
STG International Ltd., Chennai, India
 - Taught a 4 week introductory Java programming course.

Honors and Awards

- Indiana University Graduate Assistantship (*2000*)

Research and Software Projects

- Investigating the use of discrete runtime activities during the lifecycle of workflows to collect data and workflow provenance for scientific workflows and datasets derived from them. The **Karma Provenance Framework** assembles workflow activities as notifications to recreate and query the data and workflow provenance for dynamic workflows running in the NSF funded Linked Environments for Atmospheric Discovery (LEAD) meteorology and education project involving 8 research and academic organizations.

- Conducting exploratory research on the role of provenance in estimating the **quality of derived data** in scientific collaborations through metrics evaluated by statistical and machine learning methods. Also exploring uses of provenance in assisting workflow composition and resource adaptation, and long term data preservation.
- Building a **Data Movement and Naming Service** for the LEAD. The service supports different back-end storage mechanisms such as Distributed Replication Service (DRS), Amazon S3, and THREDDS Data Repository, and interfaces with name resolution provided by the Replica Location Service (RLS) from the Globus Toolkit.
- Member of the data thrust group that designed the **LEAD XML metadata schema** for representing spatio-temporal and intrinsic attributes of earth sciences data products in the LEAD project. Actively interacted with meteorologists to gather requirements and authored the first prototype schema.
- Designed and developed the **Resource Catalog**, an information service for managing organization-wide XML metadata about distributed Grid resources in the LEAD grid. The service acts as a registry for recording the WSDL and application description metadata on user services, and as a meta-catalog for crawling and indexing external data catalogs with scientific datasets, applying incremental and adaptive techniques.
- Developed the **GSX grid services library** that implemented the Global Grid Forum OGSF v1.0 specification. Built on top of the XSUL web-services SOAP library, it provides support for lifetime management of stateful services and included fundamental services such as handle resolver service, service factory, and registry services.
- Primary developer in the **XMessages project** that built a reliable messaging framework for grid and web-service applications. A publish-subscribe mechanism is used to transmit generic XML messages through message channels, with the capability to archive messages and perform historical queries using SQL and XPath. A bridge to communicate with Java Messaging Service (JMS) clients is also available.
- Participated in the **XCAT component toolkit project**, a component based framework for distributed high performance applications based on the Common Component Architecture. Tasks included the conversion of a stand-alone linear system analyzer into XCAT C++ components.
- Involved in the **Science Portals** project, a tool for scientists to develop and manage experiments and data in a distributed grid environment. Implemented an HTTP/WebDAV API in Java with wrappers to export and import web documents from a WebDAV server.
- Developed an **online exam environment** for O'Donnell Research Lab, as part of graduate level Software Engineering course. The project included developing an MS-Access front-end to design psychological tests and manage subject information, and using ASP to deliver the exams to subjects and record their results. Formal software engineering practices were followed.
- Part of the two-member undergraduate team that simulated and studied a **distributed computing environment** for a senior year project. The framework was developed in C over a Novell Netware network, and included implementing remote procedure calls, load balancing, and task switching for solving partial differential equations.

Publications

Journals and Book Chapters

- [1] Dennis Gannon, Beth Plale, Marcus Christie, Yi Huang, Scott Jensen, Ning Liu, Suresh Marru, Sangmi Lee Pallickara, Srinath Perera, Satoshi Shirasuna, Yogesh Simmhan, Aleksander Slominski, Yiming Sun, and Nithya Vijayakumar. *High Performance Computing and Grids in Action*, chapter Building Grid Portals for e-Science: A Service Oriented Architecture. IOS Press, 2007. To Appear.
- [2] Dennis Gannon, Beth Plale, Suresh Marru, Gopi Kandaswamy, Yogesh Simmhan, and Satoshi Shirasuna. *Workflows for eScience: Scientific Workflows for Grids*, chapter Dynamic, Adaptive Workflows for Mesoscale Meteorology. Springer-Verlag, 2007.
- [3] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. Karma2: Provenance management for data driven workflows. *International Journal of Web Services Research*, Idea Group Publishing, 2008. To Appear.
- [4] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. Query capabilities of the karma provenance framework. *Concurrency and Computation: Practice and Experience*, Wiley InterScience, 2007. To Appear.
- [5] Luc Moreau, Bertram Ludascher, Ilkay Altintas, Roger S. Barga, Shawn Bowers, Steven Callahan, George Chin Jr., Ben Clifford, Shirley Cohen, Sarah Cohen-Boulakia, Susan Davidson, Ewa Deelman, Luciano Digiampietri, Ian Foster, Juliana Freire, James Frew, Joe Futrelle, Tara Gibson, Yolanda Gil, Carole Goble, Jennifer Golbeck, Paul Groth, David A. Holland, Sheng Jiang, Jihie Kim, David Koop, Ales Krenek, Timothy McPhillips, Gaurang Mehta, Simon Miles, Dominic Metzger, Steve Munroe, Jim Myers, Beth Plale, Norbert Podhorszki, Varun Ratnakar, Emanuele Santos, Carlos Scheidegger, Karen Schuchardt, Margo Seltzer, Yogesh L. Simmhan, Claudio Silva, Peter Slaughter, Eric Stephan, Robert Stevens, Daniele Turi, Huy Vo, Mike Wilde, Jun Zhao, and Yong Zhao. The first provenance challenge. *Concurrency and Computation: Practice and Experience*, Wiley InterScience, 2007. To Appear.
- [6] Yogesh Simmhan, Beth Plale, and Dennis Gannon. A survey of data provenance in e-science. *ACM SIGMOD Record*, 34(3):31–36, 2005.
- [7] Dennis Gannon, Jay Alameda, Octav Chipara, Marcus Christie, Vinayak Dukle, Liang Fang, Matthew Farellee, Geoffrey Fox, Shawn Hampton, Gopi Kandaswamy, Deepti Kodeboyina, Charlie Moad, Marlon Pierce, Beth Plale, Albert Rossi, Yogesh Simmhan, Anuraag Sarangi, Aleksander Slominski, Satoshi Shirasauna, and Thomas Thomas. Building grid portal applications from a web-service component architecture. *Proceedings of the IEEE*, 93(3):551–563, March 2005.
- [8] Dennis Gannon, Randall Bramley, Geoffrey Fox, Shava Smallen, Al Rossi, Rachana Ananthakrishnan, Felipe Bertrand, Kenneth Chiu, Matt Farrellee, Madhusudhan Govindaraju, Sriram Krishnan, Lavanya Ramakrishnan, Yogesh Simmhan, Aleksander Slominski, Yu Ma, Caroline Olariu, and Nicolas Rey-Cenvaz. Programming the grid: Distributed software components, p2p and grid web services for scientific applications. *Cluster Computing*, Springer, 5(3):325–336, 2002.

- [9] Sriram Krishnan, Randall Bramley, Dennis Gannon, Rachana Ananthakrishnan, Madhusudhan Govindaraju, Aleksander Slominski, Yogesh Simmhan, Jay Alameda, Richard Alkire, Timothy Drews, and Eric Webb. The xcat science portal. *Scientific Programming, IOS Press*, 10(4):303–317, 2002.

Refereed Conferences and Workshops

- [1] Yogesh L. Simmhan, Sangmi Lee Pallickara, Nithya N. Vijayakumar, and Beth Plale. Data management in dynamic environment-driven computational science. In *IFIP Working Conference on Grid-Based Problem Solving Environments (WoCo9), to appear as Springer-Verlag Lecture Notes in Computer Science (LNCS)*, 2006.
- [2] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. A framework for collecting provenance in data-centric scientific workflows. In *IEEE International Conference on Web Services (ICWS)*, 2006. (18% acceptance).
- [3] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. Towards a quality model for effective data selection in laboratories. In *IEEE International Workshop on Workflow and Data Flow for Scientific Applications (SciFlow) in conjunction with ICDE*, 2006.
- [4] Yogesh L. Simmhan, Beth Plale, Dennis Gannon, and Suresh Marru. Performance evaluation of the karma provenance framework for scientific workflows. In *GGF International Provenance and Annotation Workshop (IPAW) & Springer-Verlag Lecture Notes in Computer Science (LNCS)*, volume 4145, 2006.
- [5] Dennis Gannon, Beth Plale, Marcus Christie, Liang Fang, Yi Huang, Scott Jensen, Gopi Kandaswamy, Suresh Marru, Sangmi Lee Pallickara, Satoshi Shirasuna, Yogesh Simmhan, Aleksander Slominski, and Yiming Sun. Service oriented architectures for science gateways on grid systems. In *International Conference on Service Oriented Computing (ICSOC)*, 2005.
- [6] Dennis Gannon, Sriram Krishnan, Liang Fang, Gopi Kandaswamy, Yogesh Simmhan, and Aleksander Slominski. On building parallel & grid applications: Component technology and distributed services. In *IEEE Challenges of Large Applications in Distributed Environments (CLADE)*, page 44, 2004.
- [7] Dennis Gannon, Marcus Christie, Octav Chipara, Liang Fang, Matthew Farrellee, Gopi Kandaswamy, Wei Lu, Beth Plale, Aleksander Slominski, Anuraag Sarangi, and Yogesh L. Simmhan. Building grid services for user portals. In *GGF Workshop on Designing and Building Grid Services, Chicago*, 2003.

Invited Papers and Technical Reports

- [1] Lavanya Ramakrishnan, Yogesh Simmhan, and Beth Plale. Realization of dynamically adaptive weather analysis and forecasting in lead. In *Dynamic Data Driven Applications Systems Workshop (DDDAS) in conjunction with ICCS*, 2007. Invited.

- [2] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. Resource catalog: An information service for community resources in LEAD. Technical Report 002, Linked Environments for Atmospheric Discovery, 2006.
- [3] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. A survey of data provenance techniques. Technical Report 612, Computer Science Department, Indiana University, 2005.
- [4] Aleksander Slominski, Yogesh Simmhan, Albert Louis Rossi, Matthew Farrellee, and Dennis Gannon. Xevents/xmessages: Application events and messaging framework for grid. Technical report, Extreme! Computing Lab, Indiana University, 2002.

Talks and Posters

- Karma2 Interoperability Challenge, Y. L. Simmhan, *Talk & Demo, Second Provenance Challenge Workshop, June, 2007*
- Effective Scientific Data Management through Provenance Collection, Y. L. Simmhan, *Talk, Microsoft Research Lab, June, 2007*
- Runtime Provenance Collection for Dynamic Information Integration and Mining, Y. L. Simmhan, *Talk, DIALOGUE Workshop, March, 2007*
- Tracking Provenance of Scientific Data in Dynamic Adaptive Workflows, Y. L. Simmhan, *Talk, IBM T. J. Watson Research Lab, January, 2007*
- Provenance in e-Science, Y. L. Simmhan, *Lecture, Tools and Technology for Computational Science graduate level course, Indiana University, Fall 2006*
- The Karma Provenance Framework v2.0, Y. L. Simmhan, *Talk, Provenance Challenge Workshop, GridWorld/GGF18 Conference, September, 2006*
- Leveraging Provenance to Facilitate e-Science, Y. L. Simmhan, *Talk, Research and Academic Computing, Indiana University, July, 2006*
- LEAD: The Linked Environments for Atmospheric Discovery, M. Christie and Y. L. Simmhan, *Science Gateways Tutorial, TeraGrid Conference, June, 2006*
- Towards a Quality Model for Effective Data Selection in Collaboratories, Y. L. Simmhan, *Talk, Indiana University CS System Seminar, Spring, 2006*
- Personal Metadata and Provenance Tracking in LEAD, Y. L. Simmhan, B. Plale, D. Gannon, R. Ramachandran, *Poster, National Forum for Geoscience Information Technology (FGIT) Meeting, October, 2005*
- Portal Resources and Service Authorization, Y. L. Simmhan, L. Fang, D. Gannon, B. Plale, *Poster, LEAD NSF Visit, July, 2005*
- Challenges in Designing a Metadata Schema for a Complex Cyber-infrastructure Project, R. Ramachandran, H. Conover, M. McEniry, S. Tanner, B. Plale, Y. Simmhan, S. Jensen, D. Lindholm, A. Wilson, J. Alameda, *Poster, LEAD NSF Visit, July, 2005*
- Information to Knowledge: Using myLEAD to Offload Mundane Tasks, N. Liu, B. Plale, S. Lee, S. Jensen, Y. Sun, Y. Simmhan, *Poster, LEAD NSF Visit, July, 2005*
- Metadata, Ontologies, and Provenance: Towards Extended Forms of Data Management, B. Plale, Y. L. Simmhan, *Talk, Networks and Complex Systems Seminar, Indiana University, Spring, 2005*
- Survey of Data Provenance Techniques, Y. L. Simmhan, *Talk, CS System Seminar, Indiana University, Spring, 2005*

- End-to-End Performance Management of DB2 Replication using Enterprise Work Load Manager (eWLM), Y. L. Simmhan, V. Gogate, I. Narang, *Talk and Poster, IBM Almaden Research Center, August 2004*
- Introduction to Grid Services, Y. L. Simmhan, *Lecture, Distributed Systems graduate course, Indiana University, Spring 2004*
- Vamana: A Scalable Registry and Discovery Service for the Grid, Y. L. Simmhan, *Talk, CS System Seminar, Indiana University, Spring, 2004*
- DB2 Support for Business Process Integration, Y. L. Simmhan, J. Myllymaki, B. Reinwald, *Talk and Poster, IBM Almaden Research Center, August 2003*
- XMessages: Building Reliable, Persistent Messaging Middleware for the Grid, Y. L. Simmhan, *Talk, CS System Seminar, Indiana University, Fall 2002*

Grants and Contracts

- “Toolkit for Provenance Collection, Publishing, and Experience Reuse”, with B. Plale and D. Leake. National Science Foundation. (*Submitted, January 2007*) .

Affiliations and Professional Service

- Member, Institute of Electrical and Electronics Engineers (IEEE) (*2000, 2005 - Present*)
- Member, Iota Nu Phi Honors Society for Informatics (*2006 - Present*)
- Member, Computer Society of India (*1999 - 2000*)
- Program Committee member, IEEE International Workshop on Scientific Workflows (SWF) (*2007*)
- Publication reviewer, HPDC (*2002*), CCGrid (*2002, 2007*), HPC Asia (*2004*), SAC (*2005, 2006*), and CCPE Journal (*2007*)
- Organization Committee member, Computer Science and Informatics Graduate Research Poster Session, Indiana University (*2007*)
- Grant reviewer, Faculty Research Support Program, Indiana University, Bloomington IN (*2006*)
- Student Volunteer, IEEE Supercomputing conference, Denver, CO (*November 2001*)

Community Service

- Organization Committee Co-Chair, Students for Bhopal 2nd Annual Conference, Bloomington IN (*September 2006*)
- Member, Academics Workgroup, Students for Bhopal (*September 2006 - Present*)
- President and Founding Member, Association for India’s Development, Bloomington Chapter (*2003 - 2005*)
- Social Chair, Computer Science Graduate Student Organization, Indiana University (*2002 - 2003*)
- Secretary, IU Cricket Club, Indiana University (*2001 - 2002*)

References

- Beth Plale (*plale@cs.indiana.edu*)
Associate Professor, Computer Science Department, Indiana University
- Dennis Gannon (*gannon@cs.indiana.edu*)
Professor, Computer Science Department, Indiana University
- Luc Moreau (*L.Moreau@ecs.soton.ac.uk*)
Professor, School of Electronics and Computer Science, University of Southampton