Data volumes in high-performance computing applications continue to increase. DOE researchers, for example, are now routinely aiming at peta- to exascale data environments, with higher resolution simulations and larger variable sets. Providing useful storage, index and selection capabilities for such data will continue to pose challenges for application and system software developers for the foreseeable future. I seek a comprehensive approach to addressing these challenges by developing:

- better tools for describing data and actions on data,
- better schemes for fine-grain control of data consistency and availability, and
- efficient, scalable system services that expose these tools and schemes to application developers.

My research centers on the role of metadata in such large-data situations. Metadata description and representation is becoming increasingly important. Novel approaches to the construction and dissemination of metadata can relieve real-time I/O bandwidth constraints while supporting existing codes. Heterogeneity of application components (codes, data, machines) is another area where intelligent, integrated metadata management can provide immediate benefits. Finally, collaboration between researchers requires common vocabularies and mechanisms for resource description, exclusion/mediation, and interoperability.

My early research investigated how the right metadata could be leveraged to provide mechanisms for protection of streaming data. Protection here refers to the ability of data producers to dynamically and differentially restrict which data items in a stream are available to consumers. My dissertation introduced *dynamic differential data protection* (D3P), defining a data access model based on familiar object- and capability-based protection mechanisms. An implementation of D3P in middleware combines intrinsic metadata (type information) with user-supplied policy statements to customize the data received by consumers. D3P can be used in high-performance and pervasive computing domains to enforce security policy by scoping or scrubbing sensitive data at run-time [2, 3], to customize data for performance reasons [7], and to provide audit facilities for run-time adaptive systems [6].

My current research reflects a continuing interest in efficient and effective strategies to describe, move, manage, and protect data at different levels in the application/operating system/hardware stack. My preferred application domain is high-performance/data-intensive computing; here I am investigating the design and implementation of software architectures for providing efficient storage in petascale scientific applications, funded by an NSF HECURA grant in collaboration with Professors Barney Maccabe and Patrick Bridges at New Mexico and Karsten Schwan at Georgia Tech. I am also collaborating with researchers at New Mexico and at Northwestern University to build a configurable open-source virtual machine monitor for HPC, also funded by NSF.

Current Efforts

Petascale Storage. Large-scale HPC applications face significant I/O challenges. This is particularly apparent for emerging classes of data-intensive HPC applications where I/O may require extensive data staging, reorganization, or conversion. Seismic exploration data processing, for example, requires multiple stages of data filtering and transformation. Each of these stages requires the loading and interpretation of hundreds of terabytes of data. These actions require not only access to massively parallel computational facilities, but also data movement at near-physical disk speeds in order to keep pace with output data production and continual checkpointing of application state. Finally, the design and successful deployment of large data-intensive applications is highly dependent on the existence of metadata (of the right kinds and in the right places); however, current I/O solutions perform the generation of such metadata in-line with application processing, reducing effective I/O bandwidth and stealing scarce processor cycles.

The key goal of this effort is to provide high-level I/O abstractions which make possible complex I/O tasks. A central abstraction is the *I/O Graph*. I/O Graphs provide a model for embedding application-specific functionality between application components. The metadata necessary to connect these components is constructed out-of-band by autonomous *metabots*, moving the performance impact of metadata maintenance out of the "fast path"

of high-end computing applications. I am a co-PI on the National Science Foundation High-End Computing University Research Activity (HECURA) grant supporting this research.

A foundation for this project is the Lightweight File System (LWFS) [4], a collaboration between the University of New Mexico and Sandia National Laboratories. LWFS investigates a "lightweight" approach to I/O, in which a core set of critical I/O functionality (access control and low-level storage) is all that is provided. Any additional storage semantics or functionality must be provided by application developers as libraries. The lightweight approach allows applications to manage their own I/O while choosing the storage semantics (and associated trade-offs) they need. This approach, by making the choices of I/O-related overheads explicit, allows the development of layered I/O subsystems which are both scalable and secure. Ongoing work on LWFS includes the design and development of lightweight transactional semantics and mechanisms, as well as application compatibility libraries.

Configurable Virtual Machine Monitor. This project builds on previous work on frameworks for development and adaptation of custom kernels for high-performance computing applications [5]. In collaboration with Northwestern University and Sandia National Laboratories, we seek to provide a high-performance, customizable, and open-source VMM for educational and application uses. The need for such a VMM has been identified by the FAST-OS Forum HPC community as well as the NSF GENI initiative. We believe that this VMM will be a fundamental tool for research in systems and architecture. For this project, I am working on adaptation and customization strategies, with an aim toward preserving customization inputs as provenance data for experiment management purposes.

Future Directions

I am particularly interested in issues surrounding data provenance, and would like to explore more fully how provenance information can be introduced and used in my chosen application domains. To do this effectively will require close collaboration with the domain researchers who are creating and using the data in question, a process already underway in the Petascale Storage project. To do this efficiently will require investigation of novel solutions in the construction of metadata, some of which are being realized in the metabot framework. I will look upward, into service specification languages and code generation tools and learn how to better represent application needs in metadata. I will also look downward into operating system, VMM, and hardware layers, to integrate with existing data typing, code mobility, and protection mechanisms as seamlessly as possible. I am interested in applying the lessons learned in high-performance application metadata design to pervasive/ubiquitous applications, where data provenance issues are also paramount. I also would like to broaden my research in these areas to applications in the healthcare domain, where questions of data protection and application interaction will only become more important as our nation's population ages.

Research Methodology and Approach

I am an experimental systems researcher, in that I build proof-of-concept systems, scale them up through iterative design-implement-test cycles, and examine them to characterize their strengths and weaknesses. When possible and appropriate, I prefer to deploy my software artifacts in collaboration with other researchers and also to make the software artifacts resulting from my research publicly available. Doing so more closely binds my work to realistic problems while allowing me to collect wider samples of usage data and feedback.

An example of how I've tried to make research artifacts generally available is the exploration of proactive approaches to information services, as realized in development of the Proactive Directory Service[1]. A prototype directory service which pushes data updates to interested parties while exhibiting good scaling characteristics, PDS continues to be used by several projects in Karsten Schwan's research group and others since its initial deployment in 2001. PDS has served as a technology base for peer-reviewed publications in venues such as IEEE/ACM Supercomputing and IEEE Cluster Computing. I am particularly pleased that PDS has been used to support research with which I am not directly involved, a statement of its usefulness and generality. Feedback from its users has, of course, been extremely valuable in guiding my research. Currently, I am developing a next generation version of this service

which will support a control plane for petascale storage architectures, and which I hope to develop and deploy in the same manner.

Another illustration of my system-building approach is found in our NSF-supported research on high-performance storage systems[8]. The applications targeted by this work run in massively parallel computational environments with thousands of cores, such as the Red Storm supercomputer at Sandia National Laboratories or the Jaguar machine at Oak Ridge National Laboratory. Obtaining access to these machines is problematic, since they are almost always fully dedicated to production code runs. While smaller exemplar versions of these machines exist, other obstacles such as graduate student nationality exist. In order to make better research progress, we built a proof-of-concept system in the Scalable Systems Lab at UNM which closely resembles the partitioned architecture used by our targeted computers. This system comprises a compute cluster, a storage cluster with parallel file system software installed, and a high-performance Infiniband interconnect. During this effort, our graduate students have had the opportunity to immerse themselves in partitioned architecture concepts and have gained valuable experience in building such a system from the ground up. Our system is capable of running small-scale problem sets using the same research codes used at SNL and ORNL. We are in the process of increasing the computational capability of this system, and are using it as the basis of collaborations with other faculty at UNM in the areas of image processing and data mining.

References

[1] Fabian Bustamante, Patrick Widener, and Karsten Schwan. Scalable directory services using proactivity. In *Proc. Supercomputing 2002*, Baltimore, Maryland, November 2002.

[2] Jiantao Kong, Ivan Ganev, Karsten Schwan, and Patrick Widener. Cameracast: Flexible access to remote video sensors. In *Proc. Fourteenth Annual Multimedia Computing and Networking Conference (MMCN'07)*, San Jose, California, January 2007.

[3] Jiantao Kong, Karsten Schwan, and Patrick Widener. Protected data paths: Delivering sensitive data via untrusted proxies. In *Proc. 2006 International Conference on Privacy, Security and Trust (PST 2006)*, Ontario, Canada, October 2006.

[4] Ron A. Oldfield, Arthur B. Maccabe, Sarala Arunagiri, Todd Kordenbrock, Rolf Riesen, Lee Ward, and Patrick Widener. Lightweight I/O for Scientific Applications. In *Proc. 2006 IEEE Conference on Cluster Computing*, Barcelona, Spain, September 2006.

[5] Rolf Riesen, Ron Brightwell, Patrick Bridges, Trammell Hudson, Arthur Maccabe, Patrick Widener, and Kurt Ferreira. Designing and implementing lightweight kernels for capability computing. *Concurrency and Computation: Practice and Experience*, 21(6):793–817, April 2009.

[6] Patrick Widener. Reverb: Middleware support for distributed application forensics. In *Proc. IEEE Workshop on Challenges for Large Distributed Environments*, Research Triangle Park, North Carolina, July 2005.

[7] Patrick Widener, Karsten Schwan, and Fabian Bustamante. Differential data protection in dynamic distributed applications. In *Proc. Annual Computer Security Applications Conference*, Las Vegas, Nevada, December 2003.

[8] Patrick M. Widener, Matthew Wolf, Hasan Abbasi, Scott McManus, Mary Payne, Patrick G. Bridges, and Karsten Schwan. Exploiting Latent I/O Asynchrony in Petascale Science Applications. In *Proceedings of the 2nd International Workshop on Parallel Programming Models and Systems Software for High-End Computing (P2S2)*, Vienna, Austria, September 2009. Held in conjunction with ICPP 2009. To appear.