Research Statement

My research interests cover a wide variety of topics in the intersection of sensor networks, databases, and distributed systems. The distinguishing aspect of my research is that I seek efficient and theoretically sound techniques to qualitatively enhance the robustness of large-scale distributed systems and I validate the techniques by implementing them in real systems. In my thesis work, I have developed new algorithms and techniques for distributed sensing services, which exploit the unique properties and requirements of target systems to provide substantially higher availability than possible with state-of-the-art techniques. Many of my key research contributions have been incorporated into a 500+ node deployment of an infrastructure monitoring service that I developed. Other researchers have used my software systems to implement publicly available services.

Current and Previous Research

The availability of low cost sensing devices and the ubiquity of network connectivity provide the opportunity to build Internet-scale *sensing services* on the information derived from live sensor feeds. An example of such services is an Ocean Monitor service [3, 7] that uses cameras and other sensors deployed in coastal regions in order to study interesting oceanographic events (e.g., riptides, sandbar formation). Prior to my work, there were no suitable generic software tools to address different aspects of building a sensing service: sensor feed processing, distributed query processing, service deployment, load balancing, fault tolerance, etc. This made authoring and deploying sensing services an onerous task, as each service author needed to address all the above mentioned aspects.

To address this, in collaboration with the researchers at Intel Research Pittsburgh, I have designed and implemented IrisNet (Internet-Scale Resource-Intensive Sensor Network Services) [2, 6]. IrisNet is the first general-purpose shared infrastructure tailored for easily developing and deploying *hybrid sensing services*, built upon powerful sensors (e.g., webcams) or embedded sensors (e.g., motes). One of the key requirements of such services is robustness against node failures and network outages. This is because the services are mostly-unattended, and more importantly, many sensing services are most useful precisely during the events that threaten system availability. For example, an ocean storm may threaten to bring the sensors and the network down, but it also generates ocean phenomena of great interest to oceanographers.

This motivates the core theme of my dissertation: What are the practical mechanisms to make sensing services highly robust? Since IrisNet focuses on building hybrid sensing services, both the wireless and the WAN components of the services need to be made robust. However, these two components are very different in nature and thus should be addressed separately.

Robustness in Wireless Sensor Networks. For wireless sensor networks, I have addressed the problem of robustly computing aggregates (e.g., the *average* temperature reported by the sensors). Traditional approaches for computing aggregates use in-network aggregation over a spanning tree rooted at the base station. However, a tree, being fragile against node- and communication-failures, gives inaccurate answers in practice. For example, under a message loss rate typical in real sensor deployments, the inaccuracy can be as high as 75%. One way to make the routing robust is to employ redundancy, by using multi-path routing for example. However, using such redundancy with traditional in-network aggregation approaches would introduce double-counting because sensor readings and partial results would be sent along multiple paths. This concern with double-counting led the researchers to stick with the tree topology despite its inaccuracies. In other words, in traditional in-network aggregation approaches, routing is dictated by the requirements of the in-network aggregation techniques.

I have proposed *Synopsis Diffusion* [10] to decouple aggregation and routing so that these can be optimized independently. Synopsis Diffusion achieves topology-independence through the use of *order- and duplicate-insensitive (ODI) synopses*. ODI synopses, a special class of synopses used in traditional data-streams research, eliminate double-counting and summarize the intermediate results during in-network aggregation.

This enables the use of a robust aggregation topology. I have shown that Synopsis Diffusion can make the aggregation process significantly more robust against typical node- and communication-failures than traditional approaches, without additional energy overheads. For example, under a typical message loss rate, Synopsis Diffusion can improve the accuracy of aggregation by around 85%. I have also provided novel synopsis diffusion algorithms for (approximately) computing a number of useful aggregates, and surprisingly simple methods to check the correctness and the approximation errors of any Synopsis Diffusion algorithm.

In an extension to this work, I have proposed, formalized, and provided algorithms to construct the *Tributary-Delta* aggregation scheme [8], a novel approach that combines the advantages of Synopsis Diffusion and traditional tree-based approaches by running them simultaneously in different parts of the network.

Robustness in Wide-area Sensor Networks. For the wide-area components of the hybrid sensing services, I investigated mechanisms to tolerate large-scale correlated failures that are common in today's Internet [12]. Evaluation of traditional robustness techniques like replication and quorum systems make several assumptions (e.g., independent failures) which, as my experience shows [13], do not hold in the current Internet. My research is the first to extensively study correlated failures in large-scale computer systems. I proposed a simple model that captures these correlated failures and used the model for analytical and experimental studies. The results show that correlated failures significantly hurt the effectiveness of traditional robustness mechanisms that assume independent failures. Additionally, correlation results in significantly diminishing availability returns for existing replication techniques.

To mitigate these adverse effects, I incorporated two components into IrisNet: a replication design based on an enhanced version of a recently-proposed quorum system [14] and a load-balancing design based on a novel database fragmentation algorithm. The replication design improves availability at the cost of a small probability of returning stale data, which is tolerable in many sensing services (e.g., oceanographers using the Ocean Monitor service are willing to accept stale data with a small probability, given that it provides service availability even when there are large failures). The load-balancing design exploits unique properties of typical sensing service workloads to significantly reduce the fragmentation time (e.g., to a few minutes compared to several hours with the existing techniques). This enables sensing services to quickly adapt to overload. Evaluation shows that my replication and load-balancing designs are effective in mitigating the negative effects of failure correlation, reducing unavailability by orders of magnitude.

Other IrisNet Research. To enable the vision of IrisNet, I have incorporated the following novel techniques within the IrisNet architecture. I have designed APIs for easily programming common tasks (e.g., data collection) of sensing services and efficient mechanisms for implementing the APIs. To achieve scalability, the APIs enable service-specific processing of sensor data near their sources and computation-sharing across services [11]. In collaboration with other researchers, I have devised mechanisms to automatically and transparently distribute the sensor data among a large number of machines and to process user queries, given in a standard XML query processing language, on that data [4, 5].

To validate the design of IrisNet, I have conducted two live deployments. First, I have collaborated with a group of oceanographers at Oregon State University who are using IrisNet to build the aforementioned Ocean Monitor service [3, 7]. Second, I have used IrisNet to develop IrisLog [1, 9], an infrastructure monitoring service, and deployed it on PlanetLab, an infrastructure currently consisting of 500+ nodes. The service has been operational and publicly available since November 2003. These live deployments have re-confirmed the importance of providing system support for robustness and showed the feasibility and effectiveness of my techniques.

Future Research Agenda

I strongly believe that the key drivers of my research (sensing services, availability of large distributed systems, etc.) are going to be increasingly important over the next decade. To illustrate that, I here present a few examples describing my future research agenda.

Highly Available Distributed Systems. The current trend of systems research suggests an increasing number of large-scale distributed system deployments in near future. At the same time, large-scale failures due to worms and attacks, software bugs, operator mistakes, etc., are becoming unavoidable. As a result, computer system availability has become as important as system performance. However, although we can often devise techniques to build high performance systems, we often lack sufficient experience to build highly available (with "five nines" of availability, for example) large distributed systems. To this end, my future research agenda includes the goal of understanding how to build highly available large distributed systems.

In my opinion, the main difficulties in achieving this goal are attributed to the following two facts. First, availability research has so far received little attention compared to performance research. We do not yet have a good understanding of many aspects of failures in real-world distributed systems. For example, we do not yet adequately know the failure patterns in large systems, we do not fully understand the heterogeneity of the failures of different system components, etc. Moreover, we do not understand the impact of system evolution (e.g. due to growth, upgrades, viruses, attacks, etc.) on these properties. We must understand these issues to devise techniques to mask failures. Second, the existing methodology to evaluate the availability of distributed systems is not sufficient. For example, there is no realistic availability workload or benchmark (similar to performance benchmarks used in file systems, for example) that we can use to satisfactorily evaluate new availability techniques. Addressing these deficiencies is extremely challenging since it requires studying large real-world systems for a long period of time.

In the near term, I plan to address some of these issues by collecting and studying failure traces of large distributed systems, analyzing the traces by using tools like statistical learning theory, investigating new robustness techniques (e.g., exploiting heterogeneity of the system components), and evaluating them by implementation in real systems.

Hybrid Sensing and Actuation Systems. I believe that distributed sensing and actuation systems, like IrisNet, will be increasingly common in near future. I envision a *world-wide sensor web* in which users can query and manipulate, as a single unit, thousands or even millions of widely distributed, heterogenous sensors and actuators. Building such complex systems requires addressing a large number of systems and algorithmic issues, only a small number of which I have addressed in my research. There remains many important but mostly-unexplored issues including the synergy between wide-area distributed systems and traditional mote-like embedded sensor systems (e.g., what is the right architecture of the gateway that connects these two systems?), distributed triggers and actuators (e.g., how to efficiently coordinate multiple actuators or to resolve conflicting actuation?), heterogeneity (e.g., given a distributed task, how does the system break it down to run within the resource constraints of different components?), sensing and actuation cost (e.g., how can the system enable users to specify the cost of a task?), privacy (e.g., how can the system protect sensitive information from the unauthorized users?), etc. Understanding these issues, crucial for the sensor web to be as useful and easily accessible as the WWW, will be a big challenge.

Another area where sensing and actuation will emerge as useful primitives is general distributed systems. Large systems will increasingly employ *software sensors* and *software actuators* that the systems themselves or the human operators can use to monitor the status or to control the behavior of different components of the systems. I believe that ideas from Synopsis Diffusion can be useful in this context. However, this would require addressing additional challenges including supporting distributed triggers, providing well-defined semantics of answers under system dynamics, finding novel aggregation algorithms, etc. We still do not know how to correctly and efficiently compute many useful aggregates (e.g., finding the distribution of the occurrences of events, computing isobars) in the Synopsis Diffusion framework. The data-streams community has explored how to compute many such aggregates, but only in duplicate-sensitive or centralized ways. I plan to leverage my theoretical results to guide algorithm design, by extending existing data-streaming algorithms or by devising novel algorithms for the Synopsis Diffusion framework.

Exploring Other Distributed Systems. The techniques I have developed in my research exploit the fact that typical sensing services can tolerate some inconsistency and approximate answers. Looking further out, there are opportunities to apply some of the lessons from my research to other domains. For example, P2P applications are expected to tolerate small data inconsistency or to accept approximate answers since, as shown by other researchers, achieving strict consistency or exact answers under the high dynamics of a P2P system would incur prohibitively large overheads. In the near-term, I plan to investigate how effective my techniques are in a typical P2P system. However, the rapid node join and leave rates of a typical P2P system would have significant side-effects on the convergence and overheads of the techniques. Addressing them effectively will be challenging.

In summary, the intersection of sensor networks, databases, and distributed systems is a rich area for future research. I hope to use my experience to address many important but unexplored issues in this intersection. I strongly believe that my research will be increasingly important in near future, and it will have a significant impact on future distributed systems.

References

[1] IrisLog: A distributed syslog. http://www.intel-iris.net/irislog.php.

[2] IrisNet: Internet-scale resource-intensive sensor network services. http://www.intel-iris.net/.

[3] The Argus Program. http://cil-www.oce.orst.edu:8080/.

[4] Chen, S., Gibbons, P. B., and Nath, S. Database-centric programming for wide-area sensor systems. Submitted for publication, 2004.

[5] Deshpande, A., Nath, S., Gibbons, P. B., and Seshan, S. Cache-and-query for wide area sensor databases. In *ACM SIGMOD* (2003).

[6] Gibbons, P. B., Karp, B., Ke, Y., Nath, S., and Seshan, S. Irisnet: An architecture for world-wide sensor web. *IEEE Pervasive Computing 2*, 4 (2003).

[7] Holman, R., Stanely, J., and Ozkan-Haller, T. Applying video sensor networks to nearshore environment monitoring. *IEEE Pervasive Computing 2*, 4 (2003).

[8] Manjhi, A., Nath, S., and Gibbons, P. B. Tributaries and deltas: Efficient and robust aggregation in sensor network streams. Submitted for publication.

[9] Nath, S., Deshpande, A., Ke, Y., Gibbons, P. B., Karp, B., and Seshan, S. Irisnet: An architecture for internet-scale sensing services. Demo and abstract in *Proceedings of Very Large Databases (VLDB)*, 2003.

[10] Nath, S., Gibbons, P. B., Anderson, Z., and Seshan, S. Synopsis diffusion for robust aggregation in sensor networks. In *ACM SenSys* (2004).

[11] Nath, S., Ke, Y., Gibbons, P. B., Karp, B., and Seshan, S. A distributed filtering architecture for multemedia sensors. In *First Workshop on Broadband Advanced Sensor Networks (BaseNets)* (2004).

[12] Nath, S., Yu, H., Gibbons, P. B., and Seshan, S. Tolerating correlated failures in wide-area monitoring services. Submitted for publication.

[13] Yalagandula, P., Nath, S., Yu, H., Gibbons, P. B., and Seshan, S. Beyond availability: Towards a deeper understanding of machine failure characteristics in large distributed systems. In *First Workshop on Real, Large Distributed Systems (WORLDS)* (2004).

[14] Yu, H. Signed quorum systems. In *23rd ACM Symposium on Principles of Distributed Computing (PODC)* (2004).