Research Statement for David Andersen

I enjoy conducting networking and systems research in a way that combines tactics from my backgrounds in both computing and the experimental sciences. In my research, I seek to create useful software systems, deploy them in real-world settings, and perform extensive experiments to characterize and explain their behavior and performance. Because modern computer systems operate in an ecosystem (such as the Internet), not a vacuum, experimental data obtained in the process of measuring deployed systems reveals important information about the underlying environment as well.

The ultimate goal of systems research, however, is the production of better systems. To this end, creating working solutions to real problems provides both great incentive and a necessary sanity check on the results of systems research. Much of my future work will retain this dual theme of experimentation and systems-building.

A major focus throughout my research has been on implementing and evaluating systems that perform well during failures and outages. Individual components of the Internet are well-engineered and often highly reliable, but the global ensemble of users, ISPs, access links, and software fails more often than we would like. I have worked on two distributed systems, RON (Resilient Overlay Networks) [3] and RAN (Resilient Access Networks) [2], which provide end-to-end mechanisms to discover and recover from Internet faults on fast time scales. RON operates by constantly measuring the paths between a small set of cooperating computers and sending data indirectly via a third computer if the direct path is unusable. RAN sends multiple session initiation packets and uses them as probes to determine a working path. In this manner, RAN provides resiliency to legacy protocols like HTTP and DNS.

Both RON and RAN have been deployed in real settings, allowing me to conduct numerous experiments. My results show that the simple approaches work well at overcoming outages: In a 20-node Internet experiment, RON avoided up to 50% of the outages between communicating hosts. In its 8 months of deployment as a Web proxy in MIT’s CSAIL, RAN has successfully masked about half of the incidents in which servers were unreachable, and eliminated 75% of certain pathological delays. The original RON research opened up considerable discussion about its scalability and fairness, which inspired many follow-on research projects, including my own work on RAN. In addition, at least one commercial product uses RON’s techniques to provide highly reliable distributed storage.

These reliability challenges extend beyond the Internet: complex software systems and locally distributed systems often present failures that are dauntingly hard to debug and prevent. Like systemically over-provisioned networks, today’s computer systems often have excess power and storage. RON and RAN provide a way to use excess network capacity to improve resiliency, and I believe that a similar approach may work for other systems. I would like to examine systems from the perspective of reliability and dependability, applying conceptually simple (but perhaps more resource-intensive) techniques that can greatly improve their reliability. Approaches that improve service availability may apply in these systems to deal with component, as well as network, failures, creating a more unified outage masking architecture.

One contribution of the RON project is that it treated the overlay network as a first-class network, not just as a vehicle for content distribution. This insight contributed similarly to my work on Mayday [1], in which the separation of overlay routing and network filtering improves both the efficiency and security of denial-of-service (DoS) prevention techniques. By examining these techniques in a realistic environment, Mayday also presented several practical and effective techniques against which future filtering schemes should be robust. The insights from Mayday about the strengths *and* weaknesses of overlay-based systems contributed to several on-going research efforts towards denial of service prevention techniques.

I believe that a thorough understanding of methods for securing computer systems and applying "security thinking" from the ground up is critical in the design and implementation of modern software systems. Building systems for real-world deployment results in an inevitable confrontation with security challenges. In the future, I plan further investigation of systems that, like Mayday, explore the boundaries and interactions between application security and network security. My experience with these systems suggests that there is a considerable efficiency trade-off involved in using only end-system techniques to achieve network security, but these techniques are typically easier to deploy and provide a good incentive structure for deployment. I don't know how far an end-system only approach can be pushed to improve network availability, but I believe that its deployment advantages make it an attractive subject for research.

Developing and sharing tools and techniques that facilitate further research is, I believe, an important duty of systems research. As part of the RON research, I deployed a medium-size Internet testbed of 36 nodes spread over 8 countries. The testbed proved essential to measuring the effectiveness of the RON approach, but was also a useful research artifact on its own—at least twelve other research projects used the testbed to gather data or evaluate their own systems, and it is currently in use by additional projects [4]. I've had the chance to contribute some of the knowledge gained from the RON testbed to the emerging Planetlab project, and have been involved since its planning stages with the design and implementation of the Emulab testbed. I believe that creating these testbeds and making them available to the wider research community can substantially improve the state of systems research, and I intend to work further on all three projects—preferably together.

Accompanying the RON testbed is a large centralized database that collects Internet measurements and routing updates from the testbed nodes to facilitate later analysis. With my colleagues, I've used the database to analyze the correlation between Internet outages and BGP updates [6], finding that most outages precede BGP messages by about four minutes, but that 20% of the failures could actually be *predicted* by a string of BGP messages. This study wasn't possible without data that combined both end-to-end probing data *and* backbone routing traces.

I also enjoy bringing new tools to bear on systems and networks problems. Using the BGP database, I showed that statistical clustering methods can be used on BGP traces to infer the internal topology of a remote network, even if that network blocks traceroute and other probes [5]. These clustering techniques are widely used in other fields for data exploration (I encountered them in the context of computational genomics); that they applied so well to network data suggests great potential for the application of other data-mining techniques to Internet measurement data.

These studies were feasible only *once the data was available.* A constant problem facing networking researchers is the lack of coherent, real-world data about network performance, and I believe we have barely scratched the surface of the utility of a well-organized, easily accessible repository of traffic, routing, and measurement data. One of my major research goals is to explore new, scalable architectures that facilitate the collection and aggregation of these kinds of data and make them available in an easily usable format to other researchers and to systems that can make use of the data in real-time. An approach that combines archival use with active use offers the potential of reduced overhead and duplication of effort, and the possibility of creating a long-term, high quality repository of measurements researchers desperately need.

I plan to continue exploring issues of resilient networked systems in the future. Resource-poor environments—lacking computation, bandwidth, connectivity, and financial resources—present an increased set of challenges for reliable systems. The growing popularity of both fixed and mobile wireless systems means that these challenges will soon be ever more present in deployed systems.

While RON and RAN worked well, I believe that broadening their goal somewhat can yield many more benefits. RON and RAN improved connectivity between specific computers, or between a computer and a service. In some cases, such as real-time communication, the RON approach matches the users needs. In many cases, however, users do not care about a connection to a computer, they care about access to *information*: an email message, a web page, their medical records, etc.. For these examples, a more powerful set of techniques can be used to improve users' ability to access them. Background replication, prefetching, alternate or redundant transport mechanisms, and other techniques can all be brought to bear. While all of these techniques could be used at the application layer, the third part of my future research will be exploring a middle ground that applies to a wide range of systems while providing most of the benefits of application integration. I don't know the exact form of this middle ground (streams? objects? opaque data blocks?) but I believe we can create an effective common layer for building much more resilient systems.

References

[1] David G. Andersen. Mayday: Distributed Filtering for Internet Services. In *Proc. USENIX Symposium on Internet Technologies and Systems (USITS)*, March 2003.

[2] David G. Andersen, Hari Balakrishnan, and Frans Kaashoek. Grassroots Reliability with Resilient Access Networks. To be submitted, February, 2004.

[3] David G. Andersen, Hari Balakrishnan, M. Frans Kaashoek, and Robert Morris. Resilient Overlay Networks. In *Proc. 18th ACM SOSP*, pages 131–145, Banff, Canada, October 2001.

[4] David G. Andersen, Hari Balakrishnan, M. Frans Kaashoek, and Robert Morris. Experience with an Evolving Overlay Network Testbed. *Computer Communication Review*, 33(3):13–19, July 2003.

[5] David G. Andersen, Nick Feamster, Steve Bauer, and Hari Balakrishnan. Topology Inference from BGP Routing Dynamics. In *Proc. Internet Measurement Workshop*, Marseille, France, November 2002.

[6] Nick Feamster, David Andersen, Hari Balakrishnan, and M. Frans Kaashoek. Measuring the effects of Internet path faults on reactive routing. In *Proc. ACM SIGMETRICS*, San Diego, CA, June 2003.