Research Statement

Bhaskaran Raman, Februray 2002

"The Parable of the Black Belt"

Excerpt from "Built to Last: Successful Habits of Visionary Companies", by James C. Collins and Jerry I. Porras.

A martial artist kneels before the master sensei in a ceremony to receive a hard-earned black belt. After years of relentless training, the student has finally reached a pinnacle of achievement in the discipline.

"Before granting the belt, you must pass one more test," says the sensei.

"I am ready," responds the student, expecting perhaps one final round of sparring.

"You must answer the essential question: What is the true meaning of the black belt?"

"The end of my journey," says the student. "A well-deserved reward for all my hard work."

The sensei waits for more. Clearly, he is not satisfied. Finally, the sensei speaks. "You are not yet ready for the black belt. Return in one year."

A year later, the student kneels again in front of the sensei.

"What is the true meaning of the black belt?" asks the sensei.

"A symbol of distinction and the highest achievement in our art," says the student.

The sensei says nothing for many minutes, waiting. Clearly, he is not satisfied. Finally, he speaks. "You are still not ready for the black belt. Return in one year."

A year later, the student kneels once again in front of the sensei.

And again the sensei asks: "What is the true meaning of the black belt?"

"The black belt represents the beginning -- the start of a never-ending journey of discipline, work, and the pursuit of an ever-higher standard," says the student.

"Yes. You are now ready to receive the black belt and begin your work."

Research Interests

My research interests are in networking and distributed systems. In particular, I'm interested in research issues that span these two domains. Middleware solutions for enabling end applications, and large-scale distributed architectures that span the Internet, pose interesting design challenges that require the application of techniques and principles from these two areas of research. My broad research interests also include heterogeneous network integration, mobile and wireless networking, networking protocol design, network measurements, and operating systems.

Research Methodology

The style of research that I have found to be very effective for networking/distributed system design is that of **system building**. Large-scale systems are often complex, and it is difficult to think of all the design issues to begin with. The methodology of **iterative design** is very useful in these cases: we have an initial design, which we implement, or build, and subsequently evaluate to identify design deficiencies and performance bottlenecks. An implementation often uncovers hidden assumptions -- a quick implementation, with the principle of "plan to throw one away" system design principle is very useful. We iterate on this process of design-build-evaluate to refine the system.

At Berkeley, I have had the opportunity to see as well as experience very successful **collaborative research**. Research projects often involve many faculty and their graduate students. Different faculty often bring in different skill sets and the collaboration yields more than the sum of the individual capabilities leading to overall success and quality in research. I also strongly believe in the synergy between **student projects** in graduate level courses, and research projects currently being pursued. Course projects are not only helpful for graduate students to kick-start their research, but also often open up exciting new directions in research. Such course projects provide the opportunity to take an in-depth look at specific research issues within the scope of the overall project. Encouraging **undergraduate projects** within the scope of the overall research goal can also often be helpful. Besides allowing graduate students a mentoring opportunity, something I found very useful in my graduate career, it also adds value to the project -- undergraduates can involve themselves in significant implementation efforts.

Another successful paradigm of research that I have experienced at Berkeley, in the context of the ICEBERG and SAHARA projects is that of **industrial collaboration**. This not only provides useful feedback to the direction of research, but also ensures a direct impact and quick use of research ideas in the real world.

Research Contributions

I have played a lead role in two projects: ICEBERG, and SAHARA. I have been active in the development of the architectures for these projects. My Master's thesis, the Universal Inbox project in ICEBERG, was an important piece and helped shape the ICEBERG architecture by being the driving application. The central theme of the SAHARA project is *service composition* across multiple providers. My PhD thesis work explores in detail a specific form of composition -- that of application level component services.

Thesis Work: An Architecture for Performance and Availability Constrained Service Composition in the Wide-Area Internet

Service Composition offers flexible ways to quickly enable new application functionality by putting together existing components. Unix piping is a very simple example of composition. In my thesis, I focused on the scenario where there are multiple service providers that develop, implement, deploy, and manage different *component* services at different points on the Internet. Other portal providers *compose* a set of these services to enable new applications. An example of such a model in today's Internet is the Yahoo portal making use of a third-party search engine to provide an end service to its subscribers. Such a model is likely to emerge given the importance of services in next-generation wired and wireless networks. There are several multimedia applications that can built on this framework of composition -- for instance, we built an application that was capable of reading out news to a cell-phone user, by composing a news source service with a text-to-speech conversion engine component. While past work has addressed performance and fault-tolerance issues in service composition within a single service provider cluster, composition in our scenario brings up new challenges. Since service providers are independent, the components could be spread across the wide-area

Internet. This makes it challenging to choose a *set* of service instances for a particular client session (more challenging than the problem of *single* web-mirror selection), based on current network and server performance. Further, studies by Labovitz et.al. have shown that Internet path availability can be very poor, and also inter-domain path recovery can take several minutes. Improving this availability by detecting and recovering from failures *quickly* is a challenge that I address in my work.

In my thesis, I have proposed and prototyped an architecture to address these performance and availability issues in service composition. My architecture is based on an overlay network of service execution platforms. The execution platforms form the middleware layer on which providers deploy their services; and the overlay network provides the context for exchange of network/server performance information exchange -- this enables an "optimal" choice of service instances for client sessions. Further, the overlay network also allows detection of network path failures, and the subsequent choice of alternate service instances. Some of the key design features of the architecture are: (a) the separation of service location information from that of performance/liveness information, (b) the client session is setup as a virtual-circuit in the overlay network -- this means that session recovery on failure need not depend on the propagation and stabilization of the failure information across the network, and (c) the use of soft-state to maintain the virtual circuit -- this makes it easy to implement session tear-down in a distributed fashion.

I have built a prototype of the system, and have evaluated it using an *emulation platform* -- there is an actual implementation of the algorithms and interfaces, and the wide-area network characteristics are *emulated* to study issues of time-to-session-recovery, scaling, and stability. My evaluations show that multi-media sessions can be recovered within a few seconds after wide-area network failures. This represents a significant improvement over Internet-path recovery which takes several 10s of seconds to several minutes.

The Universal Inbox: Providing Extensible Personal Mobility and Service Mobility in an Integrated Communication Network

While my PhD thesis concentrates on performance and availability issues in service composition, my Master's thesis was on the identification of functional components for heterogeneous device integration. There has been tremendous growth in the heterogeneity of communication networks (e.g., cellular, pager, wireless-IP) as well as communication devices (e.g., PDAs, two-way pagers, multi-model access devices). In this context, it is important to integrate the user's different communication devices: *service-mobility* is the feature that allows the user to access the same set of services from multiple devices; and *personal-mobility* is the concept of treating the end-user, as opposed to the end-device, as the communication end-point.

While there have been several efforts to integrate services across heterogeneous devices, there is no underlying architecture for such an integration. My solution for this is based on a middleware platform to support such integration in an *extensible* and *scalable* fashion. I identify three component functionalities required for personal and service mobility: (a) generic data type transformation, (b) user-customizable redirection of incoming communication based on preference profiles, and (c) device name mapping and translation. Data transformation provides data-type independence (e.g., translation between text and audio formats), and name translation achieves name-space independence (e.g., translation between pager name-space and cell-phone name-space). The preference profiles allows *control for the callee* -- the callee decides where to receive a call, as opposed to the caller deciding this.

The different components are realized as a distributed middleware platform. Their reuse for different services and devices enables easy extensibility. I implemented these components to provide integration across several end-points and services: GSM cellular-phone, desktop VoIP, PSTN phones, Voice-Mail, E-Mail, etc. The device data-type and name independence allowed us to enable services in new devices rapidly. In my implementation, such *extensions* required minimal effort. For instance, we had an Internet Jukebox service that streamed audio to desktop end-points. Extending this service to GSM cellular-phones, and PSTN phones, each took less than a few days of integration effort.

On the Layering of Protocol Functionality in Bluetooth Scatternets

Another area of ubiquitous computing that I have worked on is that of Personal Area Networks (PANs). The Bluetooth standard for PANs and home networks is based on radio-frequency technology, and specifies the physical and link-layer functionalities. While there has been a lot of research on *routing* in ad-hoc networks, as well as *service-discovery* in wired networks and home networks, there has been little study of the interaction between these two functionalities. In this work, which I did in collaboration with Pravin Bhagwat and Srinivasan Seshan at the IBM T.J.Watson Research Center, I studied this issue. Bluetooth scatternets are unique multi-hop networks in that they have a connection-oriented link-layer, with low-energy link-states to conserve power. This, coupled with the ad-hoc nature of networks, leads to

poor interaction between the link-layer, IP-layer, and the service-discovery layer. That is, if we take a traditional layered approach to the problem. We show through simulation that such a layered approach can lead to very poor performance, which nullifies the benefits of the power-saving link modes. The reason is that all the three layers try to achieve a level-of-indirection that is scatternet-wide, and repeat the same functionality. We thus argue for extensive cross-layer optimizations, and a single scatternet-wide level of indirection to achieve link-formation, IP-routing, as well as service-discovery.

Online Reordering of Data for Interactive Data Processing

Early in my graduate career, I also worked on interactive query processing in databases. Queries on large databases may take several minutes to hours. This is unsuited for interactive data analysis or viewing. In this work, done in collaboration with Vijayshankar Raman and Prof. Joseph Hellerstein, we looked at how data can be reordered dynamically to process the data of interest to the user. The user specifies preferences for which subset of data s/he is interested in, and this input is used in deciding which data to process first. Data is pre-fetched from a source (database, or spreadsheet, or live network feed), and spooled at a temporary location (called a side-disk). This pre-fetched data is re-ordered before processing an aggregate query, with the input of the user's preferences. The user sees a running aggregate of the data processed, and the statistical confidence in the aggregate given the fraction of data processed. Further, the user can dynamically change the subset of data s/he is interested in. This leads to interactive query processing; and the same idea is used for interactive viewing of huge spreadsheets.

Future Research Plans

I plan to work on research issues that are drawn from a wide range of technological background. While it is very interesting to work on issues related to cutting edge technology of the developed world, it is equally challenging to identify and work on issues in *low-cost*, *accessible*, and *appropriate* technology for the developing world. I outline my plans below.

Research Relevant to Immediate Society: Appropriate Technology

I strongly believe that research in leading technical universities should be relevant to the immediate society. This creates a synergy between research and economic development and one fosters the other. There are several factors that drive the need for a different set of technologies in the developing world:

- *The need for low-cost:* In a developing country like India, personal computers and Internet connections are out of reach for the majority of the people. However, there have been at least two significant research efforts to address this very issue: the Simputer project at IISc, and the TENET project at IIT-Madras. I believe that such efforts are likely to change the face of computing and networking in the developing world, and present a new set of research challenges.

- *Lack of reliable power:* Reliable power sources are the exception, rather than the norm in the third world. This aspect of reliable computing has received little attention so far, relying on UPS systems, which add further cost to the system. Providing reliability for different kinds of applications in this context, using techniques such as client-server computing in combination with redundancy mechanisms designed to avoid correlated failures (due to power outages) present very interesting technical challenges. This would have far reaching implications in extending computing to the rural world.

- *Different usage scenarios:* The emergence of the personal computer led to the growth of the Internet, and the emergence of hand-held devices continues to drive interest in pervasive computing. In the developing world, for reasons of economics, these models are unlikely to be the driving force. I believe that a different model of computing based on sharing compute resources, in combination with low cost networking, and low cost front-ends is more appropriate. The can have a huge impact, for instance, on computer-based teaching in rural schools, and in bridging the digital divide. Further, business relationships among owners of these resources, and the users, are likely to be very different in a rural economy. This will give rise to interesting issues in protocol/software mechanisms and policies in such an environment.

I am very interested in actively identifying and addressing such research issues in collaboration with other researchers; and in meaningfully influencing the direction of development of the immediate society.

Study of Routing Protocol Scaling

Recently, there has been a lot of interest in overlay networks and the additional resilience they provide. Availability issues arise on the Internet because of poor behaviour of BGP, and large convergence times. There is ongoing research on identifying the underlying causes for these routing issues. However, the understanding of routing protocols has turned from one of small-scale simulation or test-bed studies, to one of a poorly understood, albeit pragmatic, art of setting their various parameters and choosing between multiple variations. The reason for this is not hard to find -- the original Internet emerged as a centrally managed entity, and very small in size. Small-scale parametric studies were appropriate at this time. As the Internet emerged, effective ideas like hierarchical routing, and "hiding what happens in one part of the Internet from the rest", allowed for its successful growth. However, the Internet is now facing availability and resilience issues, and slow times to convergence.

I believe that a deeper understanding of routing protocol behaviour is required. An opportunity for this would be provided by an emulation platform -- one that can scale much better than traditional simulation-based approaches. A controlled, *large-scale* (many hundreds to thousands), design study would lead to a better understanding of routing protocol parameters, and fundamental limits to scaling and stability of routing protocols. Such a study is important before the Internet expands to encompass ubiquitous networked devices, or can support multimedia services with the reliability of the telephone network.

I plan to evolve the emulation testbed developed in my thesis for such a study. Designing such an *emulation platform for networking/systems research* is a contribution in itself, and has interesting research issues. I used a user-level emulator in my thesis. Which this gives a lot of flexibility in emulation, it lacks the performance of a kernel emulator. An interesting issue is that of the design of a *berkeley-packet-filter (bpf)* like interface. This would allow an efficient kernel implementation, while maintaining the flexibility of a user-level approach. Yet another issue in emulation platforms is that of sharing between members of the research community. The design of an emulator with *deterministic time-stretch* (e.g., 10x slower than real-time) with appropriate interfaces to user programs is an interesting issue.

Further Issues in Wide-Area Service Composition

Service Composition is a very powerful paradigm for service construction on the Internet. I have addressed a key aspect of this -- that of performance sensitive choice of service instances, and failure recovery. There are several avenues for further research. The kind of compositions that I have considered fall under the category of a *single* data flow, for a client session -- this data flow is a *path* through a set of services. However, more generically, data flow in a client session may involve a more complicated path than a string of services. For instance, it might involve a tree-like structure for the data flow, with data merging from different servers. Further, in my work, I have assumed that the *logical set* of instances in a client session remains the same. However, there are cases where allowing for this set to be dynamic may lead to better performance. For instance, composition of a compression service dynamically, based on current network performance, might improve overall network utilization. Another avenue for further exploration of service composition for mobile users. Their point of attachment to the overlay network of my architecture might change during a session, thus calling for a change in the "set" of service instances in the session.

Service Discovery Infrastructure

One of the key design aspects of my architecture for service composition was the separation of service location information from that of service liveness/performance, or network path liveness/performance. This leads to better scalability in the number of service instances. However, an aspect that is missing in the service location piece of my architecture is that of semantic queries. On the other hand, traditional approaches to service discovery allow for semantic description of services, but combine the issues of liveness/performance with location -- leading to unscalable solutions. An interesting avenue for research is to explore the possibility of using an overlay network as in my service-composition architecture, in combination with a central/replicated semantic service query engine. The former will give the separation between service location and service/network liveness, which is important for scaling; and the latter will provide searches based on rich, semantic service descriptions. The interesting research challenge here is that most of the services are likely to live outside the overlay network -- unlike in my service composition work where providers deploy their services at the nodes of the overlay network.

Bhaskaran Raman, bhaskar@cs.berkeley.edu