Keynote Talk

Multimedia Systems Research: A Retrospective OR Whatever Happened to All that QoS Research?

Harrick M. Vin
Laboratory for Advanced Systems Research (LASR)
Department of Computer Sciences
The University of Texas at Austin
http://www.cs.utexas.edu/~vin · vin@cs.utexas.edu

Over the past decade, techniques for providing Quality of Service (QoS) guarantees to applications have been the central theme for much of the research in multimedia systems. This broad category of research includes techniques for designing multimedia servers (VoD servers), network protocols and scheduling techniques for per-hop and end-to-end QoS guarantees, and QoS mechanisms for operating systems, among others. Several thousand papers have been published in these areas. Yet, few of these techniques are deployed in real systems today. Even though we understand a lot about designing systems that can provide QoS guarantees, few systems today provide any form of service differentiation or guarantees.

So, the research community should get an "A" grade for trying hard; however, I contend that the community should get a "C" or a "D" (at best) for the impact it has had on practice.

Hence, the natural questions are: what happened to all that research? Why has it had such little influence on practice? What can we learn from our experiences? It is my belief that our community should perform some introspection to answer these questions; understanding these issues will have significant impact on defining research agenda for the next decade. In what follows, I attempt to identify some of the reasons for our lack of success.

The lack of impact on practice, in my opinion, is a result of many factors. The lack of a "business case" (or charging models) is often cited as a reason for the lack of QoS support in systems today. Although this certainly is one of the reasons, it, by no means, is the only one. The following are a few other reasons to consider.

- Much of the QoS research has become victim to the "Moore's Law". Much of QoS research attempts to manage resources carefully with the intent of providing service guarantees to applications. This assumes that the resources available in the system are scarce, and must be used judiciously. This assumption is often true at the time the research is conducted; however, the community has rarely asked the question: how technology dependent is the problem (i.e., will this remain a problem in a year or two)? What level of over-provisioning will solve the problem (and how much will it cost to over-provision the system)? What is the complexity of careful resource management and how long will it take to deploy such solutions? Will the application demands keep pace with technology improvements? Although there are special-purpose applications where careful resource management is essential, for most applications, we have seen that relatively small amounts of over-provisioning eliminate the need for careful resource management. Hence, techniques for providing QoS guarantees have often lost to Moore's law!
- As a community, we have done a poor job of demonstrating the significance of the problems prior to solving them. We have often lacked applications as well as workloads for justifying or evaluating our research. We have little understanding of what QoS guarantees do applications/users really want; we have rarely argued/demonstrated that the operating regimes where QoS support is beneficial do in fact occur (or likely to occur) in practice. For instance, much of the research often assumes an operating point (e.g., high-level—90% and higher—of resource utilization); however, most service providers (network, computation or storage) rarely run their resources at these levels of utilization!
- We have paid little attention to deployment considerations for our solutions. Much of the QoS research assumes an "all-or-nothing" model for deployment—we have rarely asked the question: can these techniques be deployed and utilized incrementally? What are the benefits of partial

NOSSDAV'05, June 13–14, 2005, Stevenson, Washington, USA.
ACM 1-58113-987-X/05/0006.

deployment? Are the solutions backward compatible (how will applications developed without the knowledge of our wonderful solutions work in the new environment)? Lack of significant benefits with only partial deployments and lack of backward compatibility make it difficult for service providers to begin deployments—hence, the all-or-nothing model often fails!

- We have often solved intellectually challenging problems in designing systems that can provide QoS guarantees. However, it is often the case that actual deployments of these solutions require several "magic" numbers or help from other components (e.g., applications specifying QoS requirements). We have rarely asked the question: how difficult is it for applications/users/system administrators to deploy and use these solutions?

So, what can we learn from all this? Here are a few things to consider.

- I contend that Moore's law will help solve many of the performance-related (e.g., delay, bandwidth, loss, etc.) QoS problems for the most common applications. Careful resource management will be required only in special cases.
- Most of the systems (multimedia and others) today are too difficult to configure, manage, and use. The research community should perhaps focus on "QoS" areas such as availability, security, manageability, deployability, etc. – rather than performance issues. Making progress in these areas, however, is likely to be challenging. Today, we do not even understand how to formulate there problems precisely or evaluate solutions that address these issues!
- As a community, we should try to define benchmark applications and workloads that can provide guidelines for our explorations and evaluations. We need models for how applications, workloads, and technology will evolve—these models together will help identify fundamental problems.

In summary, it is time to evaluate our progress and develop an agenda for the future. Let's make sure that we learn from our successes and failures. Let us make "impact" as a primary criterion for developing and evaluating a research agenda. Let us develop simple, incrementally deployable solutions. Also, let us try to develop techniques that can transcend technological and applications boundaries.

This will be hard – but then who said that getting an "A" for impact will be easy?

Bio

Harrick Vin is a Professor in the Department of Computer Sciences at the University of Texas at Austin. He is the founding Director of the Distributed Multimedia Computing Laboratory and the co-Director of the Laboratory of Advanced Systems Research (LASR) at UT Austin. His research interests are in the areas of networks, operating systems, distributed systems, and multimedia systems.

He has co-authored more than 100 papers in leading journals and conferences. Harrick has been a recipient of several awards including the Faculty Fellow in Computer Sciences, Dean's Fellowship, National Science Foundation CAREER award, IBM Faculty Development Award, Fellow of the IBM Austin Center for Advanced Studies, AT&T Foundation Award, National Science Foundation Research Initiation Award, IBM Doctoral Fellowship, NCR Innovation Award, and San Diego Supercomputer Center Creative Computing Award.

He received his Ph.D. in Computer Science from the University of California at San Diego in 1993.