Scaling P2P Content Delivery Systems Reliably by Exploiting Unreliable System Resources

Hao Zhang and Kannan Ramchandran (IEEE Fellow), UC Berkeley, USA
Minghua Chen, The Chinese University of Hong Kong, China
{zhanghao, kannanr}@eecs.berkeley.edu, minghua@ie.cuhk.edu.hk

Backdrop: Internet Video Exploding

As Internet traffic continues to expand exponentially – it is predicted to be nearly four times larger in 2013 than it is in 2009 [1] – it is insightful to note that the dominant component of this explosive growth is undoubtedly rich media content, especially video data. Indeed, it is projected that Internet video alone, including live streaming, video on demand, IPTV etc. will account for over 60% of all consumer Internet traffic by 2013! And this does not even include the amount of video data exchanged through bulk-download based file-sharing applications.

This surging demand is taking its toll on existing infrastructure, such as centralized data centers and content distribution networks. Huge power consumption and high maintenance costs [2] of data centers are emerging as particular pain points in this age of environmental awareness and fiscal frugality, as content owners like YouTube [3] struggle to meet rapidly increasing demands for reliable and cost-effective streaming services. More fundamentally, this centralized infrastructure does not scale well with the rapid growth in demand. As a result, high profile video streaming failures, such as the webcast mess [4] of MSNBC's democratic presidential debate and the Oprah show web crash [5], are increasingly inevitable events.

P2P Video Content Delivery: To the Rescue

These limitations have led to a fundamental revisiting of how to deploy reliable and scalable content distribution, and triggered the inspiration of collaborative peer-to-peer (P2P) content distribution mechanisms that originates from bulk-download file-sharing applications such as BitTorrent, eDonkey etc. Indeed, P2P file-sharing has accounted for more than 60% of the Internet traffic over the past decade [1] and the technology has now evolved to accommodate video streaming applications that have stringent real-time constraints. Examples of successful P2P content delivery deployment include PPLive, PPStream and UUSee [8][9]. These commercial applications can support a large amount of demand, e.g., of up to 1,000 TV channels at an average streaming rate of 400kbps with more than 150,000 users per channel in peak time [8][9]. In 2009, a company called Octoshape used grid-cast [6], essentially a P2P-based video content distribution framework, and successfully helped CNN [7] deliver Barack Obama's Presidential Inauguration and singer Michael Jackson's live video memorial, which are the largest and the second largest video live streaming events to date as of July 2009. These events supported up to 1.34 million simultaneous users worldwide [7] which most existing centralized solutions cannot sustain without breaking the bank.

While these instances of successes based on P2P frameworks are noteworthy, is this sufficient to sustain the expected future growth? Most existing P2P streaming technologies still rely on large data centers serving as "life lines" to make up the difference in streaming rate when the users cannot by themselves redistribute the content. Since the system throughput is typically capped by the *aggregate* upload bandwidth of the participating peers which are currently bottlenecked – and are likely to remain bottlenecked in the future – by *asymmetric* connection speeds [10], the success of such systems has to depend heavily on the "big brother" peer swarms with large upload bandwidth that cooperate in sharing the content. However, these peer users are volatile in nature, making quality of service (QoS) guarantees including video quality, delay and smoothness of video playback difficult to sustain in these systems, particularly when the swarm sizes are moderate, and the asymmetry in bandwidth gets acute. This problem becomes more critical with increasing appetite for fatter bit rates required for the increasing amounts of video content and higher-definition video quality desired by consumers. As a result, content providers are often obliged to maintain over-provisioned server capacities that target worst case scenarios, which can be highly wasteful, inefficient and difficult to scale economically.

Can We Do Better? Exploiting Inexpensive System Resources

This motivates us to explore new paradigms in collaborative media content distribution that target scalability, flexibility and reliability. An interesting avenue involves judicious and opportunistic "recycling" of inexpensive, abundantly available but possibly unreliable system resources. This approach contrasts that of throwing expensive, dedicated and "deterministic" centralized infrastructure at the problem. It relies instead on a redundant supply of individually unreliable and volatile resources that may be called decentralized stochastic infrastructure. The challenge is to exploit the redundancy and abundance of these stochastic and unreliable "micro-resources" to create an aggregate reliable virtual "macro-resource" that enables us to leverage the strength in numbers to sustain an inexpensive yet scalable and quantifiably reliable overall system. There are many instances of this philosophy, and we list a few illustrative cases here:

Case 1: Cheap servers with commodity hardware. Data warehousing environments face exponential data growth, but the inability of traditional database systems to economically scale and deliver high performance has made the management and analysis of such data volumes extremely costly or, in some cases, unaffordable. On the other hand, the concept of using the statistical aggregation of cheap commodity hardware to provide reliable system services has recently received wide attention, some of which are used by Google (GFS), Java (Hadoop) and Microsoft in their file systems and associated software packages [12]. The attractive feature of this approach is that expensive, hard-to-maintain centralized bulk media servers are replaced by a redundant collection of mini-servers equipped with cheap commodity hardware. Though individually much less powerful and reliable, these can be easily obtained in larger numbers at low cost, while guaranteeing QoS through redundancy. Modern media servers and data centers that are responsible for delivering large data flows with stringent real-time constraints will potentially benefit even more from this methodology, by exploiting off-the-shelf commodity hardware that scales economically. However, this is far from a "science" today and it needs a systematic attack to realize the full potential of this paradigm.

Case 2: ISP infrastructure nodes. In another scenario, Internet Service Providers (ISPs) can deploy infrastructure nodes to aid in collaborative content distribution. Examples of these infrastructure nodes include home gateways, modems, and set-top-boxes of individual households, wireless access points, or last mile routers at the residential network pipelines. These devices can be embedded with hard disks and P2P-aware devices programmed

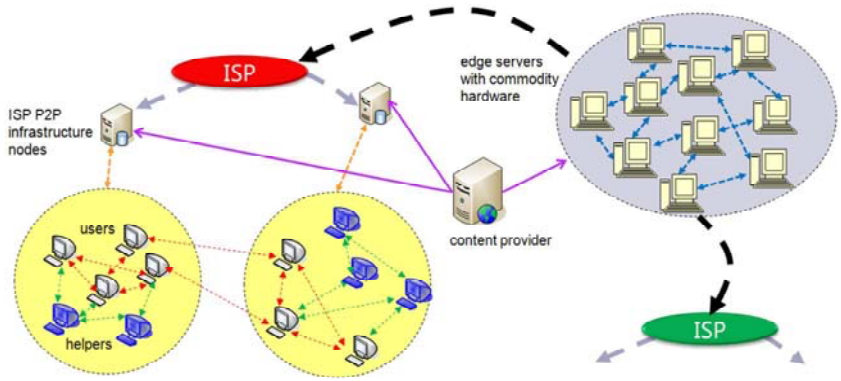

Figure 1. An example of a hybrid P2P system. Content provider pushes the content to media servers loaded with commodity hardware. ISP P2P infrastructure nodes are placed with locality awareness to save last-mile traffic. Idle Internet users contribute their resources to users of interest. Users also form a P2P network to redistribute the content among themselves.

to intelligently download and store an optimized small fraction of appropriate content, and dynamically allocate their resources to upload this content to a judiciously chosen subset of interested users. Content can be pushed into the caches of these nodes and served to users when there is demand. ISPs can potentially manage millions of these nodes as a single virtual server and take advantage of content locality and the law of large numbers using P2P infrastructure.

Case 3: Idle Internet users. Users who are merely surfing the web or checking emails or are completely idle represent a powerful collective resource for collaborative content distribution with the key differentiation with regard to classical P2P systems being that these users may not be themselves interested in the content they are helping distribute. In fact, a large number of online home pc-users are idle most of the time. As an aggregation, they often have large amount of spare storage and upload bandwidth to share. Wuala [11], an online storage system, uses similar ideas to enable users to trade in their idle resources including hard disk space and upload bandwidth which will be dynamically allocated to provide online storage service for other paying clients. In return, Wuala offers an incentive of a certain amount of free online storage to contributing idle users, depending on how much resource these users are willing to spare. This methodology and mindset not only helps create an environment for collaborative social content distribution, but also optimizes the utilization of existing resources, avoiding the waste of building excessive dedicated expensive servers.

The interpretation of "unreliable system resources" can thus be very broad. In fact, all of the abovementioned cases and any combinations thereof can be effectively adopted in practice. Figure 1 illustrates an example of such system architecture, with the goal of building a scalable and reliable system that targets real-time media content delivery and amortizes cost by exploiting inexpensive system resources, thus overcoming some of the drawbacks of centralized-only architectures.

New Opportunities for ISPs

The exploration of inexpensive system resources also brings great opportunities for ISPs. It has been widely known that the traditional P2P framework in many file-sharing applications has created an immense burden on ISPs by generating a significant amount of undesirable traffic. Since ISPs typically charge end-users only for a flat service fee for unlimited usage, P2P traffic does not generate additional revenue while consuming an enormous amount of resources. This is especially costly for ISPs when the traffic is transit among different ISP tiers. As a result, ISPs are poorly incentivized to upgrade their edge pipelines only to support the ever-increasing traffic due largely to P2P, while being marginalized by content providers' direct reaching out to consumers to cut out the "middle man".

In order to efficiently utilize system resources and maximize social welfare while obtaining reasonable shares from the booming Internet video, it may be in the ISPs' best interests to closely collaborate with both content providers and end users. Indeed, the ISPs can leverage their control of a large amount of unutilized resources – be they idle Internet users, home gateways, edge routers or high-speed fiber pipelines – thereby "outsourcing" some of the expensive maintenance components of their system cost. In a smart deployment scenario, these resources can also be dynamically allocated and optimized according to the demand. By invoking these micro-resources on a per-need basis, much of the over-provisioning of centralized architectures can be cut down, and the system will be able to scale gracefully and economically.

The broad mindset is that of leveraging the statistical aggregation effect (law of large numbers) to override micro-resource unreliability and volatility to produce a robust aggregate system. In addition, ISPs can leverage their holistic view of the Internet traffic to implement locality-aware mechanisms and avoid P2P traffic traversing long distances, saving the most expensive last mile routing. By optimizing over how much they will invest in deploying storage and P2P-aware hardware or incentivizing idle Internet users to contribute idle resources, ISPs can potentially invert the current middle man situation and boost their bottom-line profit-driven operating point derived from both content providers and end users by taking advantage of the P2P infrastructure. Both content providers and end users will also benefit from outsourcing of this task, paying the same amount of cost for better content delivery service with a guaranteed QoS. This operating point can better scale with

the boom of Internet video, and is economically friendly and socially conscious.

Challenges and Preliminary Work

The proposed framework of unreliable system resource utilization in scaling P2P content delivery brings significant opportunities as well as challenges. These unreliable system resources are typically dynamic and volatile, i.e. they are characterized by random "join and leave" behavior due to failures of cheap commodity hardware, e.g. disk failure, or due to power turnoff, or network logout, or due to just plain "free will" of these helpers who may have "something come up." How do we deploy and guarantee a system-wide robust QoS, while having to deal with the fickleness of individual micro-resources? Several important issues need to be addressed, which are all worthy of investigation, including: (1) efficient and dynamic resource deployment (who and how many resource nodes should be deployed?); (2) optimal rate allocation (how much content should the resource nodes download to be of maximum net benefit: note that they must first drain system resources before being able to help?); (3) replenishment strategies (how should the resource nodes replenish themselves to counter the effects of churn?); (4) load balancing (what allocation schedule will optimize load balancing requirements?); (5) traffic-friendly protocols (how to minimize cross-ISP traffic?); (6) system robustness (how do we architect a system which is immune to failures?); (7) content security guarantee (how do we tradeoff the benefits of collaboration with the risks of security?); and (8) incentive mechanisms (how do we motivate free idle users to contribute to system performance?)

The answers to these questions are likely to lead to highly novel system solutions. The cross-cutting nature of this research involves multimedia signal processing, communications and error-control-coding, networking protocols, optimization, and even economics and game theory. This makes for a very rich and exciting interdisciplinary problem area that is both challenging and rewarding from both academic and commercial perspectives.

In our preliminary investigations to date [3][10][13][14][15], we have exploited robust macro performance guarantees in spite of the fluctuation of micro-resources availability by exploring the law of large numbers to provide the requisite averaging effect. In particular, we have investigated how to utilize the resources of "small helper swarms" to improve the performance of P2P content delivery. Helpers are idle nodes in the P2P system who are not interested in receiving the content but have (meager) resources to spare. To design strategies that can make full use of these helpers to address the heterogeneous balance of upload and download bandwidth of common peers, careful consideration has to be paid to what and how much the helpers should download and upload before they can help system performance, and how these helpers should cooperate with peers and among themselves. A well-designed scheme will avoid making helpers drain more than they can help the system.

In [13], we studied and designed a novel scheme of utilizing these helpers for file downloading applications, and showed that these helpers will only need to download a small fixed k pieces of the data of interest consisting of N >> k pieces and transmit these pieces to peers that want the entire data. Using a fluid model with Poisson node arrival process and exponentially distributed node staying time, steady-state analysis of a simulated system have shown substantial gains of system throughput even if helpers only download a tiny fraction of the data file that does not scale with the file size. This surprising result shows even small fraction of the tiny collaboration of "helper swarms" in the social networking topology may drastically improve overall performance.

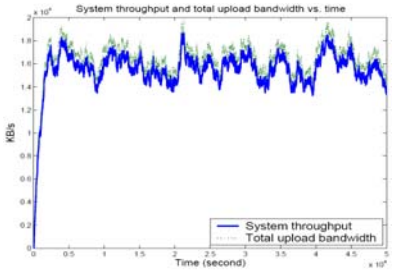

Figure 2. Throughput and total upload bandwidth of the system over time. It can be observed that throughput is kept very close to total upload bandwidth indicating that the helpers' upload bandwidths are being fully utilized.

We have also done preliminary exploration of the paradigm of P2P video live streaming. In these cases, peers have synchronized or at least loosely synchronized playback time requirements, and each packet has a specific delivery deadline. These constraints impose new challenges to the existing P2P protocols such as BitTorrent. Our work in [14] targets a helper-assisted P2P environment for live video streaming, which helps reduce the burden on content providers and bypass bottlenecks between the source and the destinations. Integrating this with distributed storage strategies [16] that are efficient to "repair" and that also leverage massive node collaboration, we have explored the architectural advantages of the helpers' use of erasure-correction parity packets in these systems. In particular, the video content owner applies a (2k, k) systematic maximum distance separable (MDS) erasure code over chunks of video data and generate k distinct parity packets for the k helper clusters. This approach enables helpers to fully utilize their upload bandwidth even if they only have 1 packet out of every k packets, and thus is robust to highly dynamic system topologies. In our simulations, we studied a P2P system with 2000 users with an average upload bandwidth of 512 kbps. Results showed that the system can sustain streaming rate of 640 kbps with 533 helpers with very little server load.

Peers	Helpers	Video bitrate	Server load (as % of total rate)
2000	0	512 kbps	2.19 %
2000	0	640 kbps	21.16 %
2000	500	640 kbps	3.20%
2000	533	640 kbps	2.40%
2000	600	640 kbps	0.70%

Table 1. Relative server load (%) v.s. different number of peers, helpers and video streaming rate.

In [3][10], we have targeted a helper-assisted P2P environment for video on demand (VoD). VoD streams are often preloaded on the server, and are available to be transmitted at users' requests. Since users can start watching the video at arbitrary time stamps, they often have asynchronous playback times. This imposes new design challenges with regard to the system architecture that satisfies users' streaming rate while maximizing helpers' utility. Using steady-state analysis, we derived the optimal system parameters as to what and how much the helpers should download and how many helpers are needed to maintain a self-sustainable system. The design is shown to yield the minimum number of required helpers and maximum utilization of helpers' resources. Simulation results showed that the proposed scheme performs very closely to the optimal stationary bound [3]. In a typical scenario of 240 users and a required theoretical minimum of 120 helpers with an average upload bandwidth of 256 kbps, a streaming rate of 384 kbps can be sustained with < 2% relative server load. Results also show that the system is robust to helper churn.

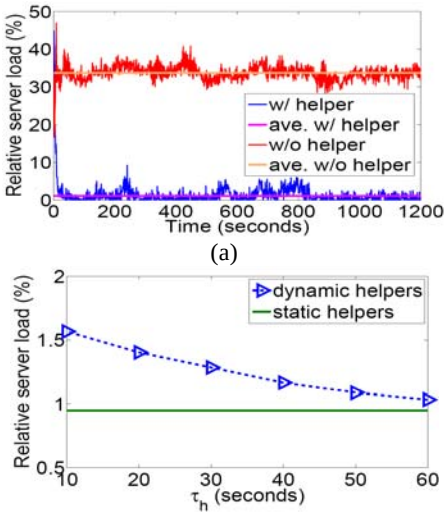

Figure 3. (a) Server load (%) v.s. simulation time. Streaming rate 384 kbps, average user/helper upload bandwidth 256 kbps, average number of users 240 and average number of helpers 120; (b) Relative server load v.s. helper's sojourn time.

In addition, Chen (one of the authors in this paper) and other co-authors Li, Chiu, and Chen have exploited the high volume of service capability in the presence of high fluctuation of micro-resources availability. They have developed queuing models for P2P service systems in which both demand and system resources come and leave randomly. The models can be applied to study the emerging P2P storage systems such as Wuala [11]. Besides answering classical questions such as system stability, the work indicates that higher server dynamics on average in fact leads to less time a work spent in the system. This suggests that there may be fundamental advantage of using unreliable, but large volume of, distributed resource against using reliable but limited centralized resource.

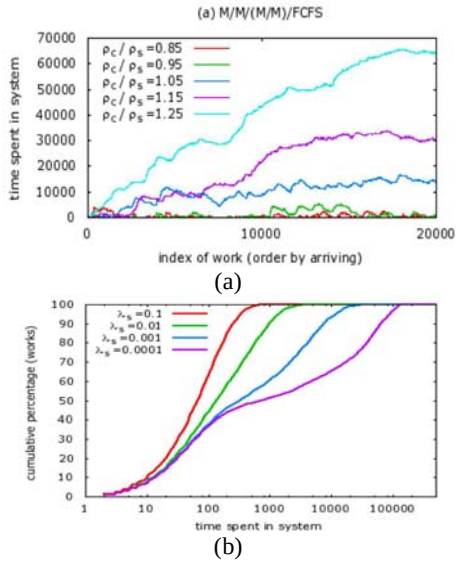

Figure 4. (a) The time work spent in the systems with different average server/client ratio; (b) Cumulative percentages for time a work spent in systems with constant average number of servers but different server churning rate. Figures extracted from [15].

All the above-mentioned preliminary investigations have shown the feasibility and efficiency of the utilization of unreliable system resources, indicating a promising future for their ability in scaling media content delivery. By exploring the strength of numbers and designing failure resilient codes, these resources can successfully provide media delivery service with guaranteed QoS, even though they are individually highly dynamic and unstable. Further investigations that include real-time optimization of rate allocation strategies to maximize system throughput, load balancing in systems with multi-video sessions with heterogeneous popularity levels, and embedded incentive mechanisms for idle Internet users to contribute their resources are highly worth of investigating in building a robust system that scales economically smoothly.

Looking Ahead

With the boom in rich media content over the Internet, the social networking aspects that leverage large scale content distribution are particularly new and interesting. We believe that the scale and scope of research should be targeted at hierarchical layers of collaboration aiming at creating novel platforms for users with common objectives and interest to interact, cooperate and share their resources. Systematic and distributed strategies that efficiently and opportunistically leverage diverse system resources are critical in obtaining solutions that meet quantitative performance metrics such as delay, bandwidth, storage capacity, and end-to-end content quality. While this vision is related in part to that of today's P2P networks and video content delivery applications, a much broader and enlightened viewpoint that looks at network interaction at extensive scales is desperately needed to keep up with the current boom of the Internet. Building systems that are scalable, reliable and low-cost will be of critical value in advancing high speed communications, high quality content distribution and efficient task dissemination.

REFERENCES

[1] Cisco Systems, Inc.. Cisco Visual Networking Index: Forecast and Methodology, 2008 – 1023. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf

[2] N. Laoutaris, P. Rodriguez, and L. Massoulie, 2008. ECHOS: edge capacity hosting overlays of nano data centers. *ACM SIGCOMM Computer. Communication Review*. 38, 1 (Jan. 2008), 51-54.

[3] H. Zhang, J. Wang, M. Chen, and K. Ramchandran. Scaling Peer-to-Peer Video-on-Demand Systems Using Helpers. *IEEE International Conference on Image Processing*. 2009.

[4] G. Daily. MSNBC makes mess of democratic debate webcast. *StreamingMedia.com*. Feb 2008.

[5] D. Kaplan. Oprah web show crashes on huge audience. *New York Post*. March 2008.

[6] http://www.octoshape.com/

[7] http://en.wikipedia.org/wiki/Octoshape

[8] C. Wu, B. Li, and S. Zhao. Exploring large-scale peer-to-peer live streaming topologies. *ACM Trans. Multimedia Comput. Commun. Appl.* 4, 3 (Aug. 2008), 1-23

[9] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross. Insight into PPLive: Measurement study of a large scale P2P IPTV system. *Workshop on Internet Protocol TV (IPTV) Services over World Wide Web*. 2006.

[10] H. Zhang, and K. Ramchandran. A Reliable Decentralized Peer-to-Peer Video-on-Demand

System Using Helpers. *Picture Coding Symposium (PCS)*, May, 2009.

[11] Wuala, the social online storage. http://www.wuala.com

[12] S. Ghemawat, H. Gobioff, and S. Leung. The Google file system. *SIGOPS Oper. Syst. Rev.* 37, 5, 29-43, Dec., 2003.

[13] J. Wang, C. Yeo, V. Prabhakaran, and K. Ramchandran. On the role of helpers in peer-to-peer file download systems: design, analysis, and simulation. *Proc. of IPTPS*, 2007.

[14] J. Wang and K. Ramchandran. Peer-to-Peer Live Multicast with Helpers. *International Conference on Image Processing (ICIP)*, 2008.

[15] T. Li, M. Chen, D. Chiu, and M. Chen. Queuing Model for Peer-to-Peer Systems. *8th International Workshop on Peer-to-Peer Systems*, 2009.

[16] A.G. Dimakis, P.G. Godfrey, M. J. Wainwright, and K. Ramchandran. The Benefits of Network Coding for Peer-to-Peer Storage Systems. *Proc. of NETCOD, San Diego*, 2007.

Hao Zhang received the B.E. degree in electronic engineering from Tsinghua University, Beijing, in 2006, and M.A. degree in Statistics and M.Sc. degree in electrical engineering and computer sciences from University of California (UC), Berkeley in 2009. He is currently working toward the Ph.D. degree in electrical engineering and computer sciences at UC Berkeley. From 2005 to 2006, he was a Research Assistant at the Center for Intelligent Image and Document Processing in Tsinghua University. He was an Engineering Intern with Cisco Systems, San Jose, CA, in the summer of 2007. Since 2007, he has been a Graduate Student Researcher with the Berkeley Audio Visual Signal Processing and Communication Systems Laboratory, UC, Berkeley. Mr. Zhang was a recipient of the U.S. Vodafone Foundation Fellowship from 2006 to 2008. He received a Best Student Paper Award at ACM MM 2008 and a Best Paper Finalist in ICASSP 2009. His research interests include image and video processing and communications, distributed source coding, computer vision, and machine learning.

Minghua Chen received his B.Eng. and M.S. degrees from the Department of Electronics Engineering at Tsinghua University in 1999 and 2001, respectively. He received his Ph.D. degree from the Department of Electrical Engineering and Computer Sciences at University of California at Berkeley in 2006. He spent one year visiting Microsoft Research Redmond as a Postdoc Researcher. He joined the Department of Information Engineering, the Chinese University of Hong Kong, in 2007, where he currently is an Assistant Professor. He received the Eli Jury award from UC Berkeley in 2007, the ICME Best Paper Award in 2009, and the IEEE Transactions on Multimedia Prize Paper Award in 2009. His current research interests include analysis and design of complex systems, distributed and stochastic network optimization and control, peer-to-peer networking, wireless networking, and network coding.

Kannan Ramchandran received his Ph.D. in EE from Columbia University, New York, in 1993. He is a Professor with the EECS Dept. at UC Berkeley, where he has been since 1999. From 1993 to 1999, he was a faculty in the ECE Dept. at the University of Illinois at Urbana-Champaign (UIUC). Prior to that, he was with AT&T Bell Laboratories from 1984 to 1990. Dr. Ramchandran received the Eli Jury Award in 1993 from Columbia University for his doctoral thesis, the National Science Foundation CAREER Award in 1997, the Office of Naval Research and Army Research Office Young Investigator Awards in 1996 and 1997 respectively, the Henry Magnusky Scholar Award at UIUC, and the Okawa Foundation Prize from the Electrical Engineering and Computer Science Department at UC Berkeley in 2001. He was the co-recipient of two Senior Best Paper Awards from the IEEE Signal Processing Society (1993 and 1997), and has been a co-author on several Best Paper and Best Student Paper awards over the past decade in leading conferences in his field. He was also the recipient of the Outstanding Teaching Award from his Department at UC Berkeley in 2009. He serves on numerous technical program committees for premier conferences in image, video, and signal processing, communications, and information theory. His current research interests include distributed signal processing and coding for ad-hoc and wireless sensor networks, robust and scalable video delivery over wireless and peer-to-peer networks, robust distributed storage, multi-user information theory, media security, and multi-scale statistical image processing and modeling..