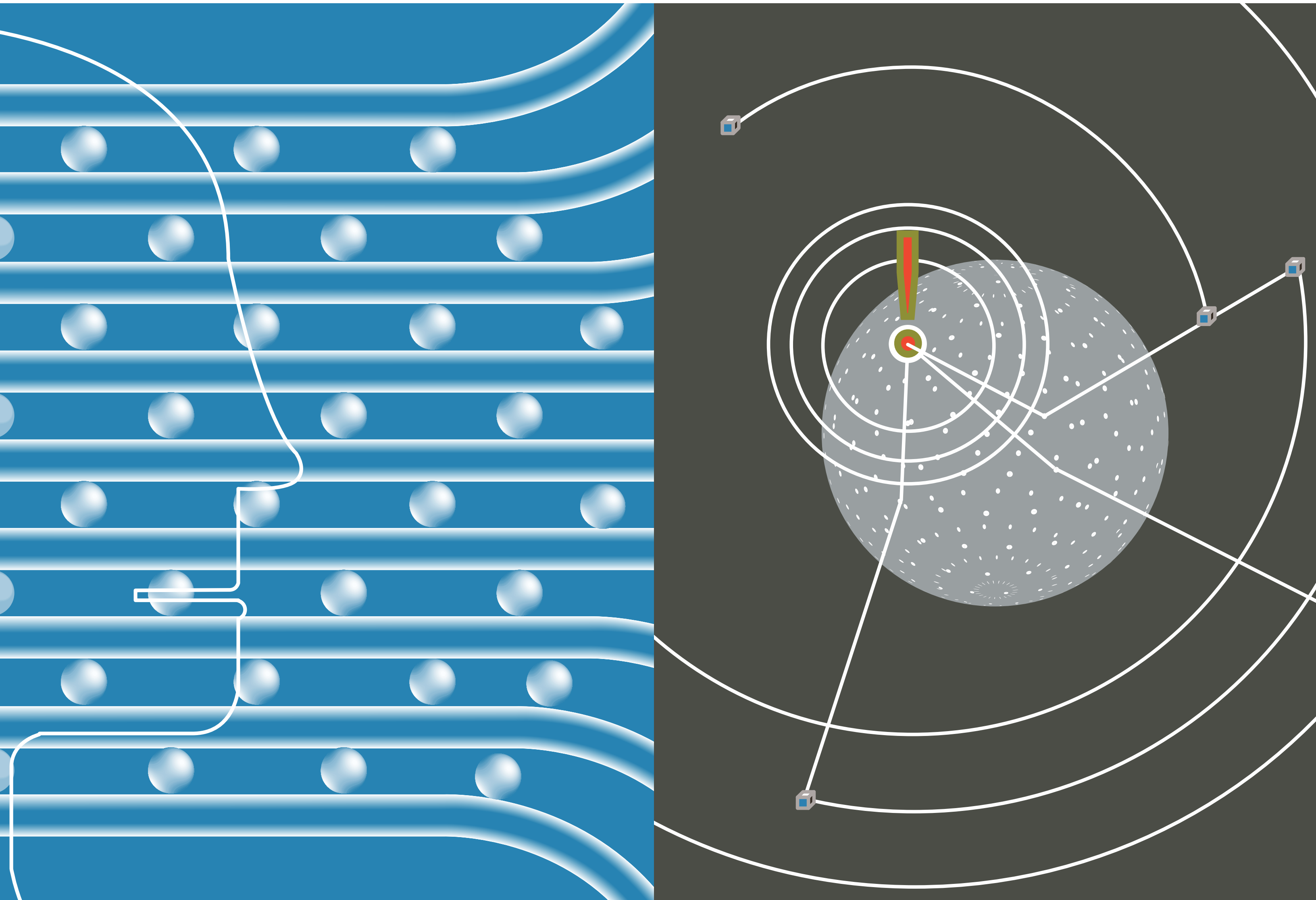
The Lee Center for Advanced Networking

CALIFORNIA INSTITUTE OF TECHNOLOGY



The Lee Center for Advanced Networking

CALIFORNIA INSTITUTE OF TECHNOLOGY

This booklet has been created to highlight and celebrate the accomplishments of the Lee Center for Advanced Networking, founded in 1999. It also serves as a thank you to Dr. David Lee whose vision and generosity made the Lee Center for Advanced Networking at the California Institute of Technology a reality.

David Lee wanted the Lee Center to support a distinctive class of research projects that are smaller and higher risk. “I wanted to give these scientists more flexibility, and allow them to explore where their research takes them.” This philosophy inspired the research conducted by the members of the Lee Center.

In regard to this booklet, I would like to thank the interdisciplinary group of faculty and researchers for their contributions; as well as Kerry Vahala, the current director of the Lee Center, and David Rutledge, the past director, for their research vision, leadership, and service. Finally, I would also like to express our gratitude to Richard (Dick) Edmiston for his guidance and advice to the Lee Center; and to Charles Wheatley for support of the yearly workshop.

Ares Rosakis
Chair, Division of Engineering and Applied Science

The Lee Center is the vision of Caltech graduate and venture capitalist, David Lee, whose David and Ellen Lee Foundation donated $10 million to establish a center for networking research over a ten-year period. Rather than support a specific single program, Lee intended the Center's researchers to have free rein "to start a lot of little projects" under the broad theme of networks. Moreover, the center, by design, would allow "people with different areas of expertise to mix together," and thereby create new synergies for research. Finally, the center would provide a powerful recruiting tool to attract the best and brightest faculty hires to Caltech.

In a time when funding for research has become ever more focused and conservative in scope, the Lee Center has created opportunities to explore those highest risk ideas that are most difficult to support, but for which the impact can be greatest. In doing so, it has had a far ranging impact on Caltech and research in general.

Its core research mission has resulted in hundreds of peer-reviewed publications spanning devices, systems and networks. In addition, the center has provided a focus to cross-fertilize traditionally disparate subject areas. Its biweekly faculty seminar lunch has brought together a diverse group of researchers; and the Lee center yearly workshop has served to summarize and regularly review progress. Through these center-sponsored interactions, new research initiatives have been inspired, such as the Information Science and Technology Initiative (IST), now located in the newly built Annenberg building. Also, research synergies have been identified, and have enabled breakthroughs such as high-frequency, CMOS power circuits for wireless systems. Moreover, several startup companies have strong ties with member groups or have derived their technology inspiration from center-funded research.

The possibility of membership in the Lee center has helped in recruiting some seven new professors; and financial support from the center has accelerated their research-program startup process. At the same time, their interaction with other members from different disciplines has provided opportunities to open new research vistas. The Lee Center has also helped to support campus projects such as the DARPA grand challenge, faculty visits from other research institutions, interactions with industry; and even projects to create local wireless hotspots on campus, long before there was a campus-wide wireless network.

Both the ten-year lifespan and support level of the Lee Center are comparable to certain National Science Foundation centers (MRSEC or STC). However, these government-funded centers tend be conservative in their scope, and focus on established research topics so as to gain both recognition and traction within the peer review system. In contrast, the Lee Center targets and literally seeds truly exploratory work. Successful projects within the Lee Center lead to much larger efforts that could become government sponsored initiatives or company startups, thereby greatly leveraging center dollars.

In summary, the Center has been a tremendous success. David Lee's idea to enable free rein "to start a lot of little projects" fills a crucial gap in how science is supported; and when combined with the breadth of the center's research scope this has enabled a range of important outcomes and synergies even beyond the support of research itself.

The impact of the Lee center is documented primarily through the online record of all center-supported publications; and is available at the Lee center website. However, faculty appointments, ongoing research initiatives, breakthroughs that have launched startup companies, and scientific collaborations both inside and outside Caltech also provide testament to its impact.

This booklet gives a brief account of the Center's accomplishments. It is organized into two parts. The first section describes research results in the words of the center members themselves; and a second section describes some, but not all, of the other areas of impact. It is our hope that David Lee's vision can serve as a model for the funding of future research at Caltech.

Kerry Vahala
Director, 2007-2010

David Rutledge
Director, 2000-2006

Natural Information Networks

A Universal Paradigm

by Jehoshua (Shuki) Bruck

Claude Shannon established the field of Information Theory in 1948. His groundbreaking work provided a quantitative characterization of information (the notions of bits and entropy), quantitative limits on information transmission (the notion of channel capacity) and methods to achieve the capacity through coding. With this new language to describe information, including, compression, and coding, the technologies for information communications and storage have flourished. The past fifteen years have presented new challenges for reasoning about information and networks, including (i) the marriage of information storage and communications, namely, the emergence of the World Wide Web and (ii) continuous connectivity between people, namely, the emergence of wireless networks. In a sense, these years saw the emergence of a connected civilization—or a natural information network. These natural information networks are not limited to the macro level, and it is also intriguing to consider micro level networks, such as biological networks, including neural networks and gene regulatory networks. Also, the emergence of natural information networks has led to studies in social science and econonomics.

My research program strongly benefited from the generous support of the Lee Center and allowed me to explore topics relating to natural networks that are not yet supported by traditional funding agencies. Examples of topics include: wireless networks and percolation theory, computation with biological networks, networks of relation and capacity of storage, and most recently, information representation in flash memories. In addition, it provided a flexible framework to attract and train the best graduate students and postdocs. In fact, two of my students supported by the Lee Center won the Wilts prize for the best PhD thesis in Electrical Engineering at Caltech (M. Franceschetti in 2003 and M. Riedel in 2004).

The Physics of Wireless Networks

Continuum percolation relates to the study of points that are distributed uniformly on the plane, where each point is the center of a disc of given radius. A natural question is: what is the minimal radius that results in an emergence of a large area of connected discs? This critical event, the emergence of a large connected area, is called a phase transition. This question has been extensively studied; in fact, in 1982 Kenneth Wilson (Caltech PhD, 1961) won the Nobel Prize for using these concepts to make calculations and predictions related to solid-state physics. Motivated by the challenges of wireless networks, especially, the notion of multi-hop routing, we considered a generalization in which a deterministic algorithm places the discs on the plane, but in such a way that each disc covers at least one point. We study the percolation properties of this new model, showing that, in the most general case, an unbounded connected component of discs does not exist. However, we identify some large families of covering algorithms, for which such an unbounded component does arise. This work introduced the tools from percolation theory to the domain of wireless networks and resulted in new methods for computing capacity, routing efficiency and reliability in wireless networks. This work was mainly done by M. Franceschetti (Caltech PhD, 2003), who is currently a tenured associate professor at UCSD.

Computation in Stochastic Biological Networks

Stochastic chemical reaction networks are among the most fundamental models used in chemistry, biochemistry, and most recently, computational biology. Traditionally, analysis has focused on mass action kinetics, where reactions are assumed to involve sufficiently many molecules that the state of the system can be accurately represented by continuous molecular concentrations with the dynamics given by deterministic differential equations. However, analyzing the kinetics of small-scale chemical processes involving a finite number of molecules, such as occurs within cells, requires stochastic dynamics that explicitly track the exact number of each molecular species. For example, over 80% of the genes in the *E. coli* chromosome are expressed at fewer than a hundred copies per cell. Further, observations and computer simulations have shown that stochastic effects resulting from these small numbers may be physiologically significant.

The power of different systems to do computation can vary greatly. It has previously been assumed that systems such as genetic regulatory networks and chemical reaction networks are much weaker than the gold standard computational systems such as Turing machines. On the other hand, we have become accustomed to proofs that even some of the simplest systems are capable of universal computation, meaning that they are in some sense equivalent in power to Turing machines. As a result, predicting their eventual behavior is impossible even in theory. Chemical reaction networks have been shown to be computationally universal when combined with some form of memory such as a polymer memory or membrane separated compartmental memory; however, researchers have previously assumed that, on their own, a finite number of species in a well-mixed medium can only perform bounded computations.

"The stochastic reaction rate foundation turns out to be the source of the computational power of chemical reaction networks."

In contrast with this historical intuition, we have shown that, in fact, such "plain" chemical reaction networks can indeed perform unbounded computation, using the concentration (number of molecules) of each species as the storage medium. Both the power and the weakness of chemical reaction network computation were also pinpointed by showing that they are as fast as a Turing machine, but require exponentially more space. This universality of chemical reaction networks derives from their probabilistic nature. If the possible reactions in a chemical system could be prioritized so that the next reaction at each step is always the one with highest priority, then universal behavior would be attainable; but, of course, chemistry does not behave in this way. However, since the reaction rates in a chemical system are influenced by the concentrations, they are somewhat under the control of the system itself and this weak form of prioritization is sufficient to let the system perform universal computation with a high probability of success. If we require that the chemical system guarantee the correct answer without fail, then the system is effectively deprived of the opportunity to use its reaction rates and in this situation, it is incapable of universal computation. Thus, the stochastic reaction rate foundation turns out to be the source of the computational power in chemical reaction networks.

This work was performed in collaboration with M. Cook (Caltech PhD, 2005), D. Soloveichik (Caltech PhD, 2008) and Professor Erik Winfree. My collaboration with Erik Winfree led to continued collaboration on the Molecular Programming Project, funded by the NSF's 2008 Expeditions in Computing Program.

Current and Future Work

Why do natural systems seem miraculous? The key reason is because we still do not know how to design systems that do what biological cells do. However, while we do not know how to design complex molecular systems, we do know how to design computers that are highly complex information systems. For example, we are capable of designing and manufacturing computer chips with billions of transistors. How did we get there? The breakthrough in digital circuit design came when we discovered a way to systematically express our ideas

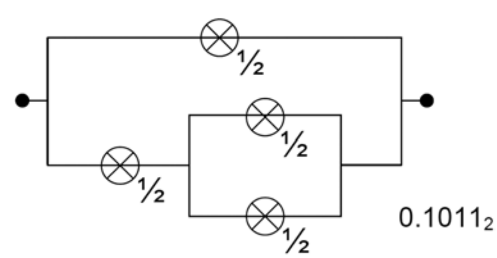


Figure 1: A stochastic network that implements the probability 11/16 with 4 probabilistic switches that are closed with probability 1/2.

using a language of symbols (calculus) that in turn could be mapped to working physical systems. Historically, two key milestones in logic design are: (i) George Boole in 1854 showing how to express syllogism (logic) with algebra (Boolean algebra), thereby, creating a calculus for logic; and (ii) Claude Shannon in 1938 showing how to implement arbitrary Boolean functions by series-parallel relay networks, hence, connecting Boolean calculus with physics.

Inspired by Shannon's work, we are now considering stochastic switching networks. These are relay networks that consist of probabilistic switches, and we study the expressive power and the reliability of these networks. In addition, we are experimenting with molecular systems, based on DNA strands, that implement stochastic networks. L E E

Jehoshua (Shuki) Bruck is the Gordon and Betty Moore Professor of Computation and Neural Systems and Electrical Engineering.

Read more at: http://paradise.caltech.edu

References

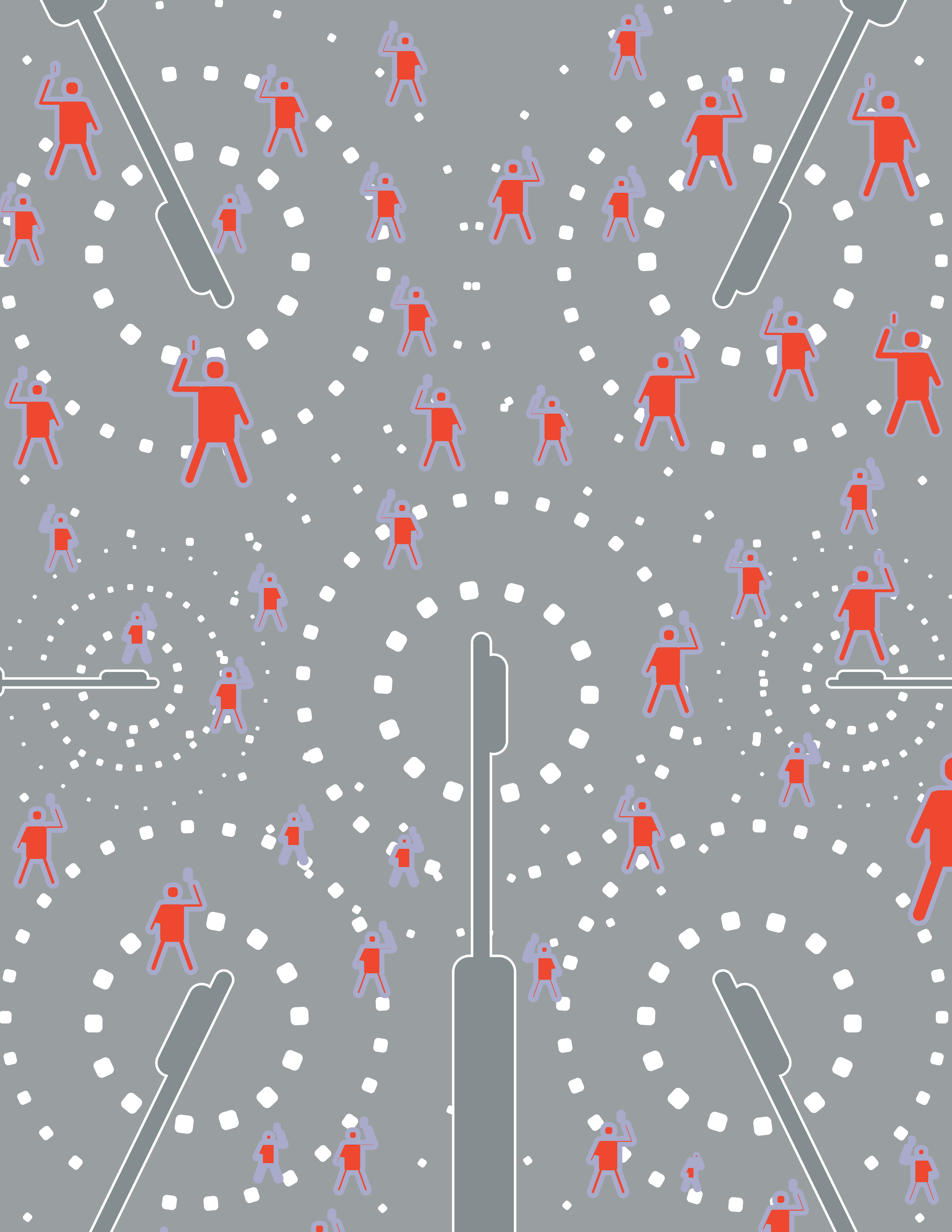
[1] A. Jiang, R. Mateescu, M. Schwartz and J. Bruck, "Rank Modulation for Flash Memories," *IEEE Trans. on Inform. Theory*, Vol. **55**, No. 6, pp. 2659-2673, June 2009.

[2] M. Schwartz and J. Bruck, "Constrained Codes as Networks of Relations," *IEEE Trans. on Inform. Theory*, Vol. **54**, No. 5, pp. 2179-2195, May 2008.

[3] D. Soloveichik, M. Cook, E. Winfree and J. Bruck, "Computation with Finite Stochastic Chemical Reaction Networks," *Natural Computing*, 2008.

[4] M. Franceschetti, J. Bruck and L. Schulman, "A Random Walk Model of Wave Propagation," *IEEE Trans. on Antennas and Propagation*, Vol. **52**, No. 5, pp. 1304-1317, May 2004. (Winner of the Schelkunoff prize for the best paper in the IEEE Transactions on Antennas and Propagation for 2005.)

[5] L. Booth, J. Bruck, M. Franceschetti and R. Meester, "Covering Algorithms, Continuum Percolation, and the Geometry of Wireless Networks," *Annals of Applied Probability*, **13**(2), pp. 722-741, May 2003.



Sense and Respond Systems

An Emerging Application of Wireless and Digital Networks

by K. Mani Chandy

Living things thrive when they sense what is going on in their environments and respond effectively. Threats and opportunities arise in many personal and business environments. Sense and respond information systems amplify the ability of individuals and organizations to respond in a timely, appropriate manner to situations that occur. Examples of such situations include: financial market turbulance; natural events such as severe weather, earthquakes, and tsunamis; and acts of terrorism. The Lee Center supported research on the theory and application of sense and respond systems. Most importantly, the Lee Center provided seed funds to do early exploration of problems when government funding for such exploration was unavailable. The results of this research have had immediate benefit to government and business; and are described in a recent book, *Event Processing: Designing IT Systems for Agile Companies*, co-authored by Mani Chandy and Roy Schulte, and published by McGraw Hill.

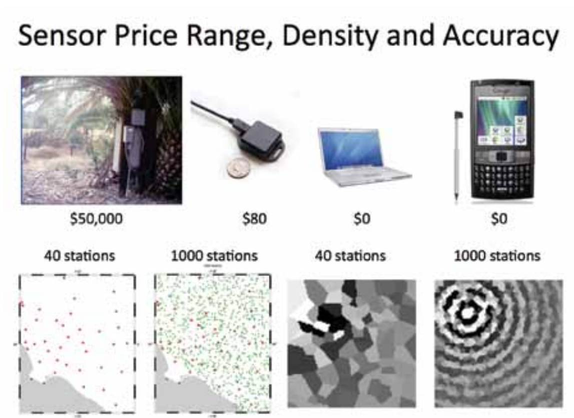


Figure 1: Earthquake Warning Systems: These systems detect early tremors in the ground due to earthquakes and warn of imminent severe shaking. The systems respond by stopping elevators, slowing down trains, and securing the electric and communication networks. A team of students and Prof. Rob Clayton in Geology, Prof. Tom Heaton in Civil Engineering, Prof. Andreas Krause in Computer Science, and Mani Chandy are carrying out this research.

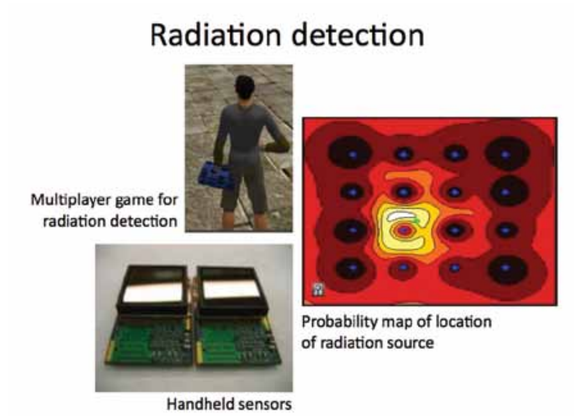


Figure 2: Radiation Detection and Interdiction Systems: These systems detect radiation from nuclear material or moved inappropriately by laboratory personnel or, as might occur in a terrorist attack. The systems respond by identifying the location of the sources of radiation and the isotopes, and then guiding security personnel. A team of Caltech students with scientists from Lawrence Livermore National Laboratory and Smiths Detection Corp is now carrying out this work.

Theory and Framework for Sense and Respond Systems

This portion of the research effort developed a common architecture for sense and respond systems. The architecture identified the key common functions—such as acquiring data, enriching data, detecting events, and responding—and then mapped the functions to enterprise applications. Businesses that want to develop support for sense and respond applications require a systematic methodology to evaluate costs, benefits and return on investment (ROI) for these applications. The Lee Center also supported research on developing such a framework.

The work first studied ways of specifying events: How does one specify what conditions to respond to? How to respond? What is normal and what is anomalous? Our work showed the promise of a compositional approach in which components dealing with different concerns are composed or "plugged together."

Lee Center research also developed theories on when and how to respond. Consider the example of a system that warns about imminent shaking from earthquakes. The data on which the warning is based is uncertain and diffuse. The system cannot wait to gather

"Living things thrive when they sense what is going on in their environments and respond effectively"

enough data to be certain because the value of a warning depends on its timeliness. How to trade-off designs that warn early but give too many false warnings with designs that give few false warnings but warn late has been studied.

Applications of Sense and Respond Systems

Besides supporting study of an architecture for sense and respond systems, the Lee center also supported research that directed these ideas towards specific applications as described below. Considerations such as use of low-cost, existing networks that can be adapted for sense and respond (see Figure 1) were considered. Much of this work was well ahead of its time when started. Although each project is now supported through government or industrial contracts, the Lee center provided the seed funding in each case.

One application concerned sensing and tracking of nuclear material, such as might occur in the case of a terrorist threat (see Figure 2). Our research dealt with integration of data from different types of sensors and data sources. For example, low-cost sensors can detect photons from radiation sources, but the sensors do not generally know the direction from which the photons came or the distance of the source. Moreover, photons are generated in an intermittent manner, and so radiation sources cannot be "seen" continuously. Lee Center funds helped to answer basic questions such as: how much time is required to detect a source in an open area with N sensors of different types? Now, the results of this research are being leveraged in contracts with the Department of Homeland Security.

A second application deals with electric grid control using modern digital technology (i.e., a smart grid). The power grid has not been an area of extensive research for the past 30 years. The area was considered staid and stable, and researchers moved on to more dynamic topics. Lee Center funds helped us to study the power grid, an area in which our group had absolutely no experience. The challenge with the smart grid is to use communication and computing technology to help manage a grid with green sources of energy, such as solar and wind, that are very dynamic. We now have a sizable contract with Southern California Edison, a major utility, to study this problem.

We expect that research on community based sense and response systems, energy, national security and medical applications will continue for several years. L E E

K. Mani Chandy is the Simon Ramo Professor and Professor of Computer Science.

Read more at: http://www.infospheres.caltech.edu

References

[1] K. Mani Chandy and W. Roy Schulte, "Event Processing: Designing IT Systems for Agile Companies," McGraw Hill, October 2009.

[2] K. Mani Chandy, Brian Go, Concetta Pilotto, Jerome White, "Verification of Distributed Systems using Local-Global Predicates" to appear in *Journal of Formal Aspects of Computer Science*, April 2010.

[3] K. Mani Chandy, "The Impact of Sense and Response Systems," *IEEE Internet Computing*, Vol. **14** No. 1, February 2010.

[4] "Networked Sensing Systems for Detecting People Carrying Radioactive Material," Proceedings Fifth International IEEE Conference on Networked Sensing Systems, June 2008.

[5] "Towards a Theory of Events," Proceedings Conference on Distributed Event Based Systems, 2008.

The Architecture of Robust, Evolvable Networks

Organization, Layering, Protocols, Optimization, and Control

by John C. Doyle

Biological systems are robust and evolvable in the face of even large changes in environment and system components, yet can simultaneously be extremely fragile to small perturbations. Such universally robust yet fragile (RYF) complexity is found wherever we look. The amazing evolution of microbes into humans (robustness of lineages on long timescales) is punctuated by mass extinctions (extreme fragility). Diabetes, obesity, cancer, and autoimmune diseases are side-effects of biological control and compensatory mechanisms so robust as to normally go unnoticed. RYF complexity is not confined to biology. The complexity of technology is exploding around us, but in ways that remain largely hidden. Modern institutions and technologies facilitate robustness and accelerate evolution, but enable catastrophes on a scale unimaginable without them (from network and market crashes to war, epidemics, and global warming). Understanding RYF means understanding architecture—the most universal, high-level, persistent elements of organization—and protocols. Protocols define how diverse modules interact, and architecture defines how sets of protocols are organized.

"robust yet fragile"

Insights into the architectural and organizational principles of networked systems can be drawn from three converging research themes. (1) With molecular biology's description of components and growing attention to systems biology, the organizational principles of biological networks are becoming increasingly apparent. Biologists are articulating richly detailed explanations of biological complexity, robustness, and evolvability that point to universal principles. (2) Advanced technology's complexity is now approaching biology's. While the components differ, there is striking convergence at the network level of architecture and the role of layering, protocols, and feedback control in structuring complex multiscale modularity. New theories of the Internet and related networking technologies have led to test and deployment of new protocols for high performance networking. (3) A new mathematical framework for the study of complex networks suggests that this apparent network-level evolutionary convergence within/between biology/technology is not accidental, but follows necessarily from the universal system requirements to be efficient, adaptive, evolvable, and robust to perturbations in their environment and component parts.

Additional Details and Selected References

The Internet is an obvious and familiar example of how a protocol-based architecture facilitates evolution and robustness. The hourglass protocol "stack" has a thin, hidden "waist" of universally shared feedback control (TCP/IP) between the visible upper (application software) and lower (hardware) layers. This allows "plug-and-play" between modules that obey shared protocols; any set of applications that "talks" TCP can run transparently and robustly on any set of hardware that "talks" IP, accelerating the evolution of TCP/IP-based networks. Diverse applications use additional protocols to exchange objects such as web files, images, and email. In the hardware layer, signals are also exchanged via shared protocols and interfaces. Indeed, recent theory proves that with suitable versions of TCP/IP and network provisioning, an otherwise arbitrary (e.g. large and heterogeneous) network has guaranteed global dynamics that robustly optimizes aggregate application utility. (Similar results hold for flocking and oscillator synchronization.) Other theory shows that overall system robustness obeys conservation laws, suggesting RYF behavior can be managed but not completely eliminated.

Within a layer these protocols can be visualized as "bowties" with large fan-ins and -outs of energy, materials, and information via a thin "knot" of universal protocols specifying carriers, building blocks, or interfaces. Other engineered examples of bowties include networks which connect energy sources to consumers via carriers and standard socket-plug interfaces, sellers to buyers via money, fibers into diverse clothing via sewing, and raw materials to assemblies via standardized building blocks for advanced manufacturing. In these and the biologic examples below, the currencies, carriers, intermediates, precursors, plugs, packets, conserved residues, and interfaces in the bowtie "knot" are highly constrained by protocols. Yet this shared universality allows robust adaptation and evolution to extreme diversity at the inputs (energy, sellers, fibers, materials, nutrients, receptors) and outputs (consumers, buyers, clothing, assemblies, products, effectors), as long as they have appropriate (and typically hidden) layers of feedback control.

The robustness and evolvability of the cell exemplify reliance on layers and protocols analogous to those in TCP/IP and other engineered architectures, but with wondrously different hardware components and more extensive and sophisticated use of feedback control. Perhaps the ultimate bowtie involves the universal, central protocols that transcribe and translate DNA to proteins via conserved codes, polymerases, and RNA intermediates. As in IP, feedback control of polymerases yields a functional network with regulated levels of macromolecules. A simple cell in a constant environment needs little more, but real cells add robustness to changing demands with a TCP-like feedback layer of allostery and post-translation modifications acting on faster time scales. These TCP/IP-like layers (the waist of a cell's hourglass) of control protocols constitute much of biochemistry texts.

"plug-and-play"

In the microbial biosphere, genes that "talk" transcription and translation protocols can move by horizontal gene transfer (HGT), thus accelerating evolution in a kind of a "bacterial internet." Furthermore, proteins newly acquired by HGT or through duplication and divergence or domain rearrangements can work effectively because they can exploit additional shared protocols. For example, a bacterial cell's application layer consists of biochemical fluxes with various additional bowtie protocols such as core metabolism (nutrients to biosynthesis and energy supply via conserved precursors and carrier-based group transfers) and signal transduction (receptors to downstream effectors via conserved residue pairs). Thus selection acting at the protocol level could evolve and preserve this variety of shared and ultimately conserved architecture, essentially "evolving evolvability".

All life and advanced technologies rely on protocol-based architectures. The evolvability of microbes and IP-based networks illustrate how dramatic, novel, dynamic changes on all scales of time and space can also be coherent, responsive, functional and adaptive, despite implementations that are largely decentralized and asynchronous. New genes and pathways, laptops and applications, even whole networks, can plug-and-play, as long as they obey protocols. Biologists can even swap gene sequences over the Internet in a kind of synthetic HGT. A related aspect of sophisticated architectures is that actuator signals (e.g. from intermediate fluxes and concentrations in core metabolism to cardiopulmonary rates and renal concentrating effects) have extremely high variability in order to keep other critical signals (e.g. from key metabolic products to blood oxygenation and pH, to core body temperature,) tightly regulated despite fluctuating supplies and demands. Such natural physiological variability can be a source of confusion if homeostasis is misinterpreted as implying that all signals, including actuators, are held nearly constant.

Typical behavior is fine-tuned with this elaborate control and thus appears boringly robust despite large internal and external perturbations and fluctuations. As a result, complexity and fragility are largely hidden, often revealed only by rare catastrophic failures. Since components come and go, control systems that reallocate network resources easily confer robustness to outright failures, whereas violations of protocols by small malicious or even random rewiring can be catastrophic. So programmed cell (or component) "death" is a common strategy to prevent local failures from cascading system-wide. The greatest fragility stemming from a reliance on protocols is that standardized interfaces and building blocks can be easily hijacked. So that which enables HGT, the web and email also aids viruses and other parasites. Large structured rearrangements can be tolerated, while small random or targeted changes that subtly violate protocols can be disastrous. Thus another source of high variability is in the external behavior of RYF systems, where typical behavior is boringly robust but rare events can be catastrophically large.

Chaos, fractals, random graphs, criticality, and related ideas from statistical physics inspire a popular and completely different view of complexity, where behaviors that are typically unpredictable and fragile "emerge" from simple and usually random interconnections among homogeneous components. The attraction is understandable because a variety of features, continually recycled in series of "new sciences," greatly simplify modeling and analysis. For example, random internal rewiring has little effect. Tuning, structure, and component heterogeneity is minimal, as is system functionality, robustness, and environmental uncertainty. Perhaps most appealing is that "modularity" reduces to superficial self-similar patterns in ensemble averages, avoiding the complexity of protocols and architecture. Unfortunately, highly evolved and architecture-based RYF complexity is utterly bewildering when viewed from this perspective, which has led to widespread and remarkably persistent errors and confusion about complex systems of all kinds. There are striking parallels with irreducible complexity and intelligent design, which has even more successfully popularized persistent errors and confusion, though more narrowly targeting biology and evolution.

Particularly persistent errors surround the origin and nature of power laws. The statistical invariance properties of Gaussians and exponentials for low variability phenomena are well-known enough that their presence is little cause for surprise. But for high variability phenomena, power laws have even richer statistical invariants, and could rightly be called "more normal than Normal." Thus their ubiquity in the abundant high variability of architecture-based RYF systems should be no surprise, and certainly no evidence on their own for any specific mechanism beyond the presence of such high variability.

One thing all can agree on is that the search for a "deep simplicity and unity" underlying natural and technological complexity remains a common goal. Fortunately, our growing need for robust, evolvable technological networks means the tools for engineering architectures and protocols are becoming more unified, scalable and accessible, and happen to include the more popular complexity views as extreme special cases within a much richer structured, organized, and dynamical perspective. These developments are bringing much-needed rigor and relevance to the study of complexity generally, including biology, but will not eliminate the need for attention to domain-specific structure and detail. Quite the contrary: both architectures and theories to study them are most successful when they facilitate rather than ignore the inclusion of domain-specific details and expertise.

John Doyle is the John G Braun Professor of Control and Dynamical Systems, Electrical Engineering, and Bioengineering.

Read more at: http://www.cds.caltech.edu/~doyle

References

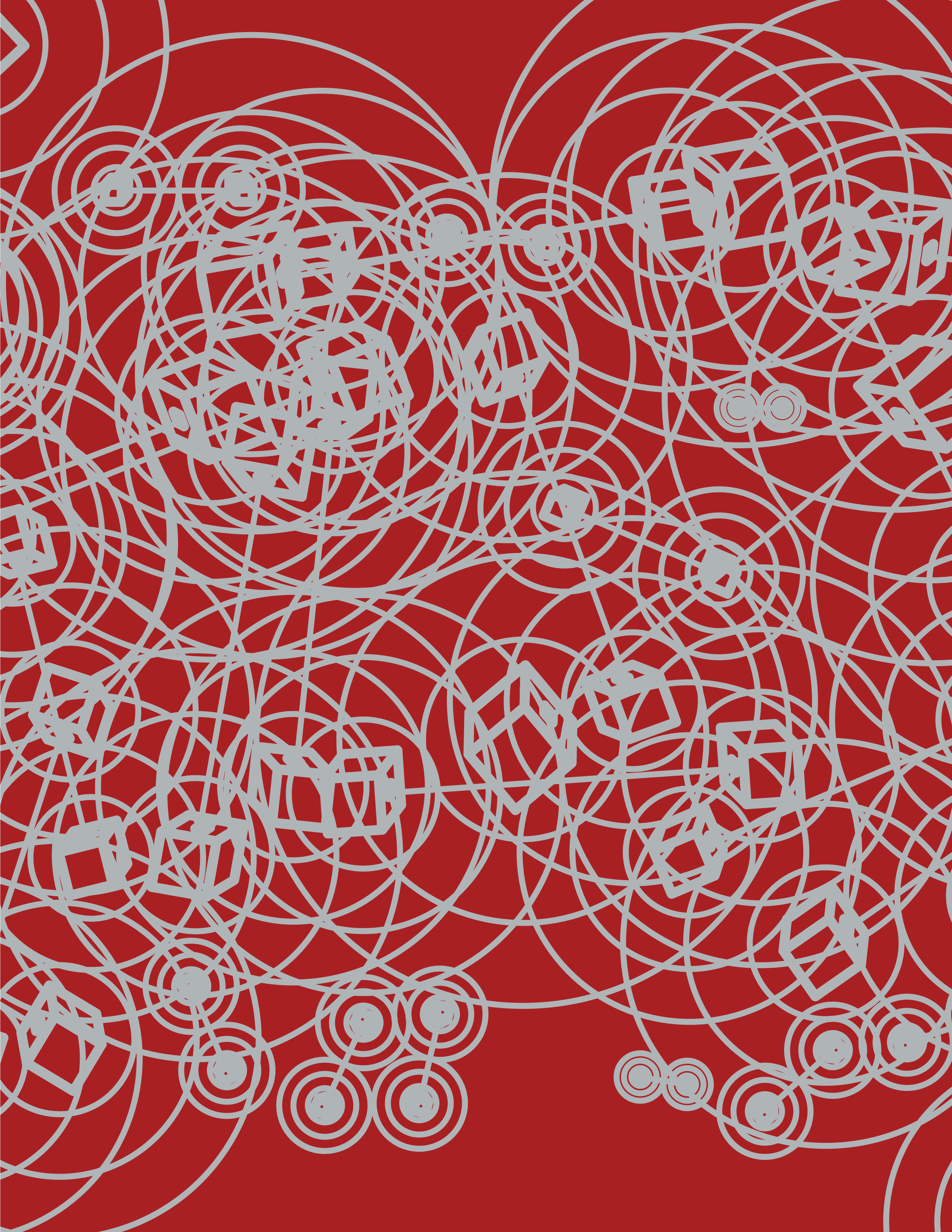
[1] H. El-Samad, H. Kurata, J.C. Doyle, C.A. Gross, and M. Khammash, "Surviving Heat Shock: Control Strategies for Robustness and Performance," *PNAS* **102**, 8 (2005).

[2] Doyle et al, (2005), "The 'Robust Yet Fragile' Nature of the Internet," *PNAS* **102**, 41, (2005).

[3] M. A. Moritz, M. E. Morais, L. A. Summerell, J. M. Carlson, and J. Doyle, "Wildfires, Complexity, and Highly Optimized Tolerance," *PNAS*, **102**, 50 (2005).

[4] H. El-Samad, A. Papachristodoulou, S. Prajna, J. Doyle, and M. Khammash, "Advanced Methods and Algorithms for Biological Networks Analysis," Proceedings of the IEEE, **94**, 4, 832-853 (2006).

[5] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering As Optimization Decomposition," Proceedings of the IEEE, **95**, 1 (2007).



Research in Two-sided Matching Markets

An Experimental Study of Stability and Compromise Market Outcomes

by Federico Echenique

Two-sided matching markets are an important form of economic organization. Labor markets are simple examples, as are many other markets with clearly delineated buyers and sellers. In some special but important cases, the markets are centrally organized. For example, in the market for medical interns in the US (among other countries), participants can opt to submit matching preferences. A potential intern may state that Hospital 1 is the most-preferred choice; followed by Hospital 2, and so on. Hospitals submit their preferences in a similar fashion. The centralized organization then uses an algorithm to determine who is matched with whom.

"we investigate the predictive power of the theory of stable matching"

We have studied centralized markets experimentally; and have conducted a collection of experimental studies—initially funded by the Lee Center—of two-sided matching markets. The research is joint with Leeat Yariv, Associate Professor of Economics at Caltech, Gabriel Katz (graduate student at Caltech) and Alistair Wilson (graduate student at NYU). In our baseline treatment, there are two groups of agents (a metaphor for firms and workers). Agents can form pairs, obtaining different payoffs from matching with different agents from the opposite group, with very few constraints on their interactions (i.e., in a decentralized fashion). We investigate the predictive power of the theory of stable matching, analyzing the type of matchings that emerge (namely, stable or not) as a function of the size of the markets and the preferences of participants. In addition, we test which stable matchings have more prominent drawing power experimentally.

We find that stable matching, the rationale behind centralizing the medical interns market, is a powerful driving force. The vast majority of markets seek to achieve a stable matching. Additionally, there are very interesting selection dynamics among stable matchings: a market will seek out a compromise, a matching that is neither too good for buyers and bad for sellers or vice versa. This selection finding is unexpected as the currently-used centralized algorithms install a matching that is best for one side and worst for the other; currently, for the interns market, it is the best for the interns. The are no known efficient algorithms for an intermediate matching, yet markets on their own seek to install one.

		Stable	O. F.	O. C.	Median	Unstable	Mkt. Stable
8×8	Unique	95%				5%	92%
	Two	94%	61%	39%		6%	91%
	Three	95%	11%	16%	73%	5%	72%
	Total	95%				5%	85%
		Stable	O. F.	O. C.	Median	Unstable	Mkt. Stable
15×15	Unique	83%				17%	50%
	Three	100%	3%	3%	93%	0%	100%
	Total	89%				11%	80%

The table above shows the main results of the first collections of experiments we have described. The labels "O.F." and "O.C" stand for the matchings that are optimal to each of the sides of the market (buyers and sellers). The "median" matching is the compromise between both sides. The two sizes are indicated on the left of the table: 8 or 15 on each side of the market. We also indicate whether there was a unique, two or three stable matchings in the experimental design. Note that there is a median stable matching only when the design has three stable matchings.

We conjecture that the intermediate matching is driven by the symmetry in market position of the two sides. However, experiments in which only one side of the market is allowed to propose a match show, surprisingly, that the intermediate matching is still prevalent.

In a separate set of experiments, we are investigating the institutions used in centralized matching markets. We want to understand what are the incentives facing the agents who participate in these in-

"the protocols mimic the mechanics of the real-world matching institutions we study"

stitutions. Our experiments are designed so that agents in a market follow a specific protocol for arriving at a match; the protocols mimic the mechanics of the real-world matching institutions we study. Our results suggest a significant volume of manipulation on the side of participants and emerging outcomes that are limited in efficiency.

Matching markets have an important role in the economy. They determine such varied outcomes as which student goes to what school, who gets which job, who marries whom, and who buys which house. Our experiments are some of the first to analyze crucial elements determining decentralized and centralized market outcomes. This new experimental platform is designed to complement a range of empirical studies of markets and will offer a controlled way to assess important factors dictating the functioning of different types of markets. We are very active in this direction, and working on many problems that remain open, such as accounting for incomplete information, restricted communication between agents, and many others. LEE

Federico Echenique is Associate Professor of Economics.

Read more at: http://www.hss.caltech.edu/~fede

References

[1] Gabriel Katz and Leeat Yariv, "An Experimental Study of Decentralized Matching,'' Caltech Social Science Working Paper [in progress].

[2] Alistair Wilson and Leeat Yariv, "An Experimental Study of Centralized Matching,'' Caltech Social Science Working Paper 1315, December 2009 [unpublished].

[3] Sang Mok Lee and Matthew Shum, "Aggregate Matching,'' Caltech Social Science Working Paper 1316, January 2010 [unpublished].

Network Equivalence

New Tools for Understanding Network Capacities

by Michelle Effros

Caltech's Lee Center has played a critical role in my group's research into the field of network information theory. Research in this area aims to understand the limits of network communication systems. One of the earliest results in the field is by Slepian and Wolf (1973); and considers a network with two transmitters describing information to a single receiver. For example, imagine two students, Alice and Bob, sitting in the same classroom. Both are taking notes, and each plans to send those notes to their friend Charlie, who is home sick with the flu. Since Alice and Bob are listening to the same lecture, their notes are likely to be similar but not identical. Each understands the material a little differently. Each misses a different portion of what is said in class. Slepian and Wolf set out to understand how long the descriptions sent to Charlie must be in order for him to learn everything contained in both sets of notes. If neither Alice nor Bob knew of the other's notes, then the total description length would be long; anything that appeared in both sets of notes would be described twice. In contrast, if Alice and Bob got together and prepared a single set of unified notes, the combined description could be much shorter. Slepian and Wolf studied the description length needed when Alice and Bob work independently. In this case, each knows that the other is also sending notes to Charlie, but neither knows what the other's notes say. Slepian and Wolf demonstrated that it is possible for Alice and Bob to independently describe their notes in a manner that allows Charlie to learn everything in both sets of notes from a pair of descriptions whose total length is no longer than the shortest description that Alice and Bob could have devised had they worked together. Thus Alice and Bob can do as well as if they worked together even though neither sees the other's notes (see Figure 1).

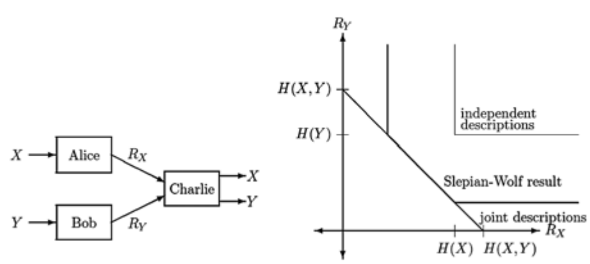


Figure 1: As shown by Slepian and Wolf, the shortest possible descriptions from Alice and Bob are far shorter than those required if neither knew the other were sending a description.

Slepian and Wolf's startling result illustrates both the power and the challenges associated with the field of network information theory. The power of the result lies both in the remarkable new understanding that it gives us about the world and in the technological advances that it enables. For example, more and more sensors are making their way into our environment. They monitor ground movement for signs of earthquakes and buildings for evidence of structural weakness; they track the movements of storms and their impact on tidal wave patterns; they help us to understand the health of our forests and oceans and air. As the number of sensors increases, the quantities of information to be gathered and the savings to be obtained by taking advantage of correlation between measurements increase as well. Advances from network information theory can help us to tackle these problems. On the other hand, the challenge illustrated by the Slepian-Wolf example is that the specific network studied is small because deriving this type of result is very difficult. In fact, almost four decades later, the gap between the complexity of the networks for which we have such complete characterizations and the complexity of the networks through which we communicate daily is growing ever larger.

With support from the Lee Center, we are pursuing methods for moving network information theory beyond such small networks. The goal is to develop rigorous, systematic tools for understanding the limits of large, complex communication networks. We wish to understand how much information these networks can carry, how to operate networks in order to enable the maximal reliable transfer of information, and how to design new networks in order to increase the amounts of information they can carry while decreasing the power and other resources required to meet those communication goals.

There are many challenges. First is the complexity of the networks themselves. In networks that rely on wireless components like mobile phones and sensors, information is transmitted through the air. Signals sent through this shared medium are overheard—to some extent—by all devices in the region. As a result, the signal that I receive on my cell phone is a very complicated mixture of the transmissions of all devices in range—even the devices with which I am not trying to communicate. The more devices communicate across a given network, the more complicated it becomes to understand how to protect information from noise and interference and to analyze how much information can be reliably delivered through the network.

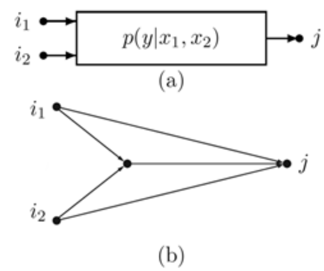


Figure 2: (a) A multiple access channel and (b) its noiseless bit-pipe model.

We have introduced a new equivalence theory for network capacity as a first step towards the development of systematic tools for network analysis. The idea is to model the behavior of wireless and noisy wireline network components in terms of noiseless wireline models. The resulting deterministic networks are easier to analyze, and a variety of tools for performing such analyses are already available. While the idea of breaking networks into simpler components and analyzing those components individually is not new, prior component models capture the performance of network components in isolation but fail to capture their full range of performance when these components are employed in larger network systems. Therefore, the key to our approach is to develop models for which we can demonstrate that no matter how the given component is used in a larger network context, the model can emulate the component's behavior. By building a library of wireless and noisy wireline component models, we can bound the communication capabilities of any network comprised of these components by the communication capabilities of a corresponding network of noiseless network components.

Examples of network components that we have modeled to date are the noisy wireline, broadcast, multiple access, and interference channels. The broadcast channel is a network component in which a single transmitter sends information to multiple receivers. The message intended for each receiver is different, but all receivers hear different noisy versions of everything sent by the transmitter. The multiple access channel is a network component that models interference (see Figure 2) wherein, multiple transmitters independently send information to a single receiver. What the receiver hears is some noisy combination of the signals sent by all transmitters. The interference channel combines the broadcast and interference effects into a single network component. In this component, each transmitter wishes to communicate with a single, distinct receiver. Since all transmitters send their signals simultaneously, each receiver hears a combination of its desired signal mixed with the signals intended for other receivers and with noise.

As this work continues, we are increasing the library of models and working to automate the process of separating complex practical networks into the components that most accurately capture network behavior as a whole. Combining the resulting algorithms with existing tools will enable analysis of increasingly complex communication networks. By understanding the limits of these network models and the strategies that achieve those limits, we hope to improve the use of current communication networks and create improved network designs in the future. LEE

Michelle Effros is Professor of Electrical Engineering.

Read more at:

http://www.ee2.caltech.edu/people/faculty/effros.html

References

[1] D. Slepian and J.K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, IT-**19**: 471-480, 1973.

[2] R. Koetter, M. Effros, and M. Médard, "On a Theory of Network Equivalence," In proceedings of the IEEE Information Theory Workshop, Volos, Greece, June 2009.

[3] R. Koetter, M. Effros, and M. Médard, "Beyond Network Equivalence," In *Allerton Annual Conference on Communications, Control, and Computing*, Monticello, IL, September 2009.

[4] M. Effros, "On capacity upper bounds for a simple family of wireless networks," In proceedings of the Information Theory and Applications Workshop, San Diego, CA, February 2010.

Low-power Data Communication Circuits for Advanced Integrated Systems

by Azita Emami

Advanced electronic systems require complex architectures that consist of many integrated circuit (IC) chips or modules. Examples of such systems are high-performance multiprocessors, servers, backplanes of internet routers, consumer electronics, and biomedical devices. There is also a growing demand for multiple processing and memory units integrated on a single chip. The high-bandwidth communication between these ICs and modules is a critical design issue. Large numbers of high-speed inputs and outputs (IOs) are required to create efficient interfaces between different processors, memory units and other modules located at varying distances from each other. With the continuous scaling of feature sizes in chip manufacturing technology, the speed of on-chip data processing as well as the level of integration will continue to scale. In order to enhance the overall performance of the system, the bandwidth of the interconnections needs to follow the same trend. To achieve this goal, scaling of the data rate per IO as well as the number of IOs per chip is necessary. While the increased switching speed of transistors allows faster transceiver electronics, the scaling of interconnect bandwidth has proven to be very difficult, and new approaches are necessary. The main challenges are: power and area limitations, properties of highly-scaled technologies, and characteristics of the signaling media.

Synchronous data transmission, shown in Figure 1, is the most common interconnection scheme in integrated systems. A precise clock signal and a driver are used to launch the data into the channel media. The channel is defined here as a complete signal path from the transmitter to the receiver, which includes the packages, Printed-Circuit-Board (PCB) traces, vias, and connectors (Figure 1a). Different modulation and coding schemes can be used for transmission of the information. The most basic and widely used scheme is binary data transmission by current or voltage amplitude modulation. The goal is to recover the transmitted bits at the receiver with a very low Bit-Error-Rate (BER) at very high speed (data rate). The information is recovered by the receiver, which samples the incoming signal using a synchronous clock.

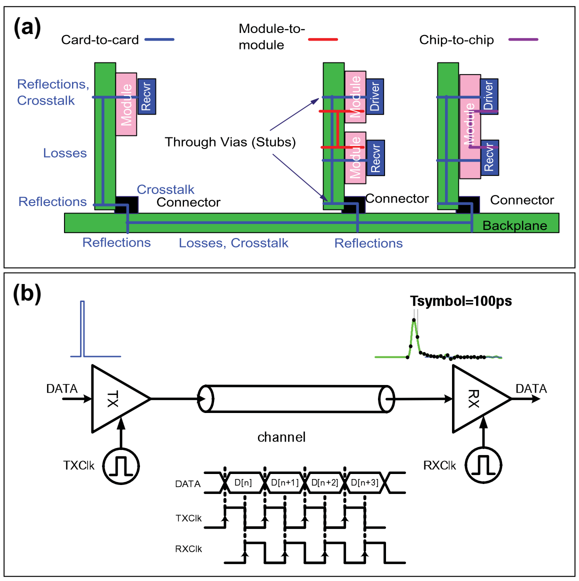


Figure 1: (a) Electrical links at different levels; (b) Synchronous data transmission.

Under support from the Lee center, we have studied new solutions for high-bandwidth chip-to-chip, module-to-module and intra-chip interconnections in highly-scaled technologies. We envision that a promising approach for data communication in advanced integrated systems will be an optimum combination of electrical, optical, and wireless signaling techniques. Accurate link modeling and understanding the fundamental limitations of different signaling methods have been among major goals of this effort.

Novel All-digital Synchronization Loop

One of the key building blocks of a synchronous data transmission system is the timing recovery loop at the receiver. Driven by power dissipation issues in high density circuits, timing and voltage margins are continually being reduced. It is therefore essential to precisely sample the signal at the time that the Signal-to-Noise-Ratio (SNR) is maximum. We designed a novel fully digital architecture that removes the need for an analog Phase-Locked-Loop (PLL). In this approach, an extra clock phase is used to continuously monitor the incoming signal and to find the location of the transitions. After the completion of each monitoring period, the functionality of the data and monitor clocks are switched. The ping-pong action between the two clock phases is the key to the effectiveness of this architecture. A 10Gb/s data receiver based on this method has been implemented in a 90nm CMOS technology, and is currently under test and measurement. The results confirm that this receiver is capable of recovering an arbitrary amount of phase shift between clock and data without use of an analog PLL[1].

Low-power Optical Transceiver

Instead of furthering the complexity of electrical links, a different approach is to change the signaling media entirely. The possibility of using optics for interconnection at short distances has been a subject of considerable research and analysis. By providing a high-capacity channel, optical signaling can potentially close the gap between the interconnect speed and on-chip data processing speed. Although optical signaling involves electrical-to-optical (EO) conversion and vice versa (OE), its large channel bandwidth can simplify the design of the transceiver electronics. The maximum data rate of an optical link is in fact limited by the performance of the optical devices and the speed of on-chip electronics. The number of optical IOs per chip is usually limited by the power consumption and area of the electronics. For parallel optical signaling at short distances, one can either use fiber-bundles, integrated waveguides or free space to send collimated beams in parallel. In all cases the cross-talk among the beams is relatively low, avoiding another problem with large numbers of electrical IOs.

A three dimensional configuration for parallel optical interconnection in free space is shown in Figure 2.

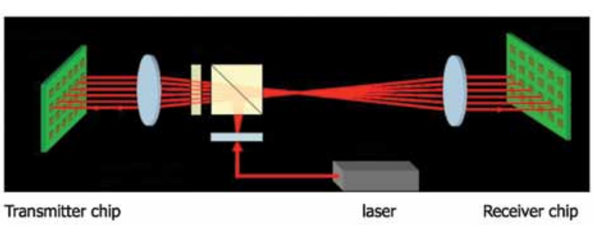


Figure 2: Parallel optical chip-to-chip interconnection over free space, with optical devices flip-chip bonded to CMOS chips.

The integration of dense two-dimensional arrays of optical devices with standard CMOS ICs has been demonstrated, and allows a huge chip-to-chip interconnection bandwidth[2,3]. In prior work, to achieve low-power and low-area optical IOs, a novel receiver front-end was designed using a double sampling/integrating technique. This approach facilitated a number of interesting scalable solutions such as parallelism, de-multiplexing, and an efficient baud-rate clock and data recovery. The transceiver operated at data rates as high as 16Gb/s and consumed less than 130mW[3]. These promising results proved the merits of optical signaling and the importance of continuing of this approach. Under support from the Lee center, we identified some of the shortcomings of the double sampling/integrating receiver and have found novel solutions for them. The integrating front-end suffers from headroom problem and needs DC-balanced signaling. In the new front-end, we are applying the double sampling technique to a resistive front-end. The added resistor converts the optically generated current signal to a voltage signal and automatically limits the voltage swing at the input node. The double sampling extends the bandwidth of the system beyond the time constant of the input node.

Efficient electrical, optical and hybrid signaling for on-chip and chip-chip networks are among the most important requirements of future computing and communication systems. We will continue to focus on understanding the fundamental limitations and developing of ultra-low-power solutions. Our effort will include optimum pulse shaping at the transmitter, novel cross-talk cancellation and equalization techniques, and further use of optics to enhance performance. LEE

Azita Emami is Assistant Professor of Electrical Engineering.

Read more at: http://www.mics.caltech.edu

References

[1] Matthew Loh, Azita Emami-Neyestanak, "All-Digital CDR for High-Speed High-Density I/O," submitted to the *2010 IEEE Symposium on VLSI Circuits*.

[2] A. Emami-Neyestanak et al. "CMOS Transceiver with Baud Rate Clock Recovery for Optical Interconnects," Proc. *IEEE VLSI Symp on Circuit*, June 2004.

[3] S. Palermo, A. Emami-Neyestanak, M. Horowitz, "A 90nm CMOS 16Gb/sTransceiver for Optical Interconnects," *IEEE Journal of Solid-State Circuits*, **43**, No. 1235-1246, pp. 889-896, May 2008.

Distributed Integrated Circuit Design for the 21st Century

by Ali Hajimiri

Ten years ago, we noticed that a significant milestone was on the horizon. It arrived a few years ago when the cut-off wavelengths of silicon transistors became smaller than the dimensions of typical chips implemented in those processes[1]. This crossover had fundamental implications in terms of the way high-frequency silicon-based circuits should and will be architectured and designed in the coming decade, making it possible to address a broad range of applications. Moreover, on account of the remarkable progress in the field of integrated circuits, it is possible to integrate more than a billion functional transistors capable of operating at mm-wave frequencies on a single silicon die. This (practically) unlimited number of high-speed transistors combined with the crossing-over of chip-size and cut-off-wavelength has produced unique opportunity that is just beginning to be harnessed.

Many new and existing applications can benefit from high-frequency, distributed circuits implemented in silicon. There are sensing and ranging applications using mm-wave frequencies. For instance, low-cost, highly integrated radars in applications such as automotive radar have received a great deal of interest in the recent years[2, 3, 4]. Automotive radars can be used in a broad range of applications, e.g., early warning and brake priming, self-parking, global traffic control, low-visibility driving aid, autonomous cruise control, and collision avoidance. On the communication side, in addition to traditional multi-gigabit data transmission, new direction-dependent modulation schemes, such as near-field direct-antenna modulation[5], can make it possible to create secure and power efficient data streams with concurrent full-rate transmission in several directions. Integrated power generation at high frequencies is another major challenge that has been helped by distributed integrated circuits. Finally, mm-wave silicon-based circuits can find applications in security, medical imaging and biochemical sensors.

To take full advantage of the opportunities offered by the die-size wavelength crossover, one must retool for a new era. The artificial borders between electromagnetics, antenna, propagation, device physics, as well as analog and digital circuit design must be removed. These levels of abstraction, originally established to partition the overall system design into tractable tasks, are no longer valid at mm-frequencies and above. This new global design approach provides a larger design space and enables architectures and topologies with better performance and new features. The Lee center supported work to developed not only this new design approach, but also its application to several new CMOS-based technologies. One of these is now described. Another is detailed later in a description of our startup company Axiom.

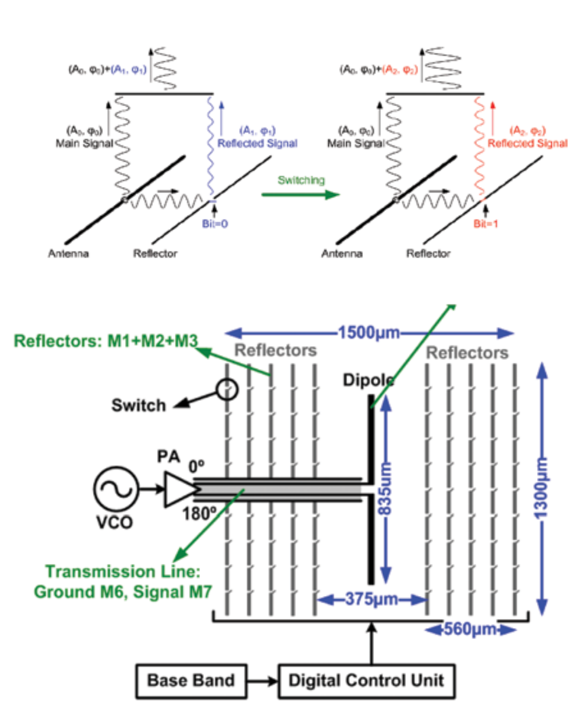


Figure 1: (a) simple one bit direct antenna modulation with a single switch. (b) 90-bit modulation at 60GHz

Near-field Digital Antenna Modulation

Conventional radio transmitters take information at a so-called baseband frequency and then up-convert and amplify that information to a stronger signal at the desired (higher) transmission frequency. This process requires upconversion mixers and perhaps more importantly a linear (or linearized) power amplifier. As the final step, the amplified signal is fed to an antenna that radiates the modulated signal in all directions, albeit with different gain factors and time delays.

The near-field digital antenna modulation technique is a fundamentally different approach to data transmission that applies the signal directly at the antenna through digital manipulation of the electromagnetic boundary conditions[5]. By combining steps at the antenna it is possible to simplify the remainder of the transmitter since up-conversion mixers and linear power amplifiers are no longer required. The antenna modulation technique can also be used to create a secure communication channel by sending the desired data only in the intended direction, while a scrambled data stream is transmitted in other directions. Moreover, it can be used to concurrently transmit two (or multiple) streams of completely independent data in different directions both at the maximum rate, thereby increasing the overall transmitted data rate significantly. Finally, the simple design enables data transmission at high rates that are not limited by the bandwidth of the up-conversion chain.

Figure 1 shows the basic principle behind the digital antenna modulation technique, where in this case a dipole antenna is driven by a continuous-wave signal of constant amplitude and phase. Also shown is a conductive metal line with comparable dimensions to the wavelength (i.e., a reflector) next to the antenna. This can be shorted or opened at some point along its length using a switch. The reflected signal interferes with the main signal radiated by the antenna in a given direction. The amplitude and phase of the reflected signal depend on the boundary conditions that the reflector imposes and can be varied in time by turning the digital switch on or off. These on/off states of the switch result in two different phases and amplitudes in the desired transmission direction. This provides a simple one-bit digital modulation without changing the output power or phase of the power amplifier driving the antenna. The amplifier therefore operates at its highest efficiency.

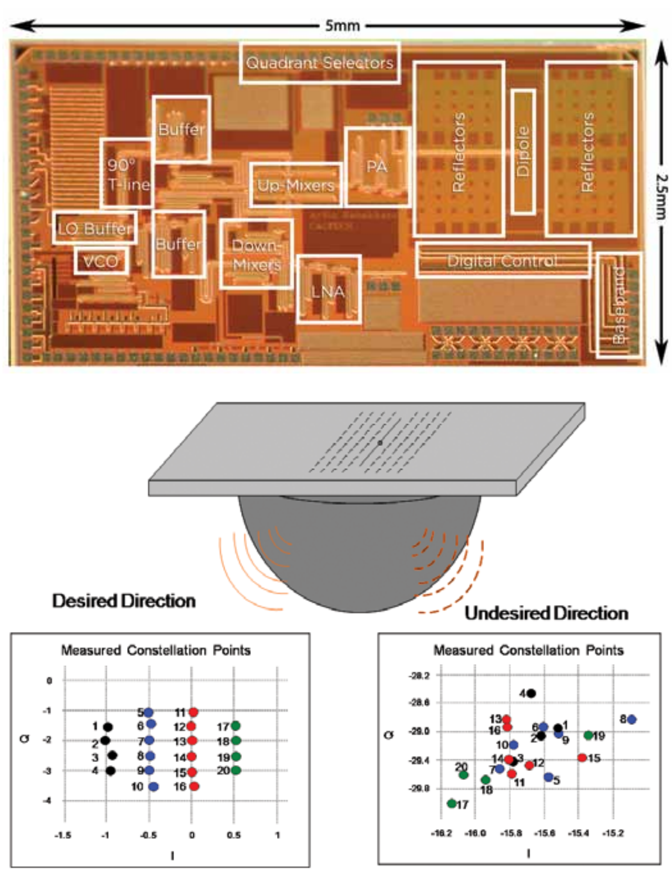


Figure 2: (a) die micrograph of a direct antenna modulation system at 60GHz implemented in silicon (b) measured signal constellations in the intended and unintended directions.

Although the single-reflector, single-switch configuration of Figure 1 provides basic binary modulation, it does so with only limited control over the direction and information content of the transmitted signal. A larger number of reflectors and switches in close proximity to the antenna makes it possible to precisely control the desired direction, and even transmit distinctly different information streams in different directions, simultaneously. This latter feature can be used to scramble the signal further in the unintended directions to implement a secure communication link by preventing an undesired eavesdropper from demodulating and recovering the signal.

The viability of this technique was demonstrated in an integrated circuit implementation[5] containing an on-chip dipole antenna with 10 reflectors. Each reflector featured 9 tuned MOS switches along its length, for a total of 90 switches and 2^{90} (~10^{27}) switching combinations. The chip and other details of the design are illustrated in Figure 2. The large number of switching combinations provides numerous ways to generate a desired point on what is called an information constellation (i.e., point in phase and amplitude) along a given direction. This, in turn, provides many additional degrees of freedom that can be used for concurrency or security. The ability to simultaneously send independent information to several directions at full rate using a single transmitter is not possible using conventional transmitter architectures.

This technology in which antenna reflectors are switched in the near field using a large number of very fast switches is a direct result of our ability to integrate, antennas, radio-frequency electronics, and digital circuits on the same substrate. This is but one example of the global co-design of electromagnetic structures with analog and digital circuitry. In practice, designing such system poses challenges not only at the architecture and circuit levels, but also with respect to the simulation tools and methodology. The Lee center seeded research on these ideas nearly 10 years ago, launching what has become a major new field within electronics. LEE

Ali Hajimiri is Thomas G. Myers Professor of Electrical Engineering.

Read more at: http://www.chic.caltech.edu

References

[1] A. Hajimiri, "The Future of High Frequency Circuit Design," Plenary Talk, European Solid-state circuits Conference, Sep 2009.

[2] A Babakhani, X Guan, A Komijani, A Natarajan, and A. Hajimiri, "A 77-GHz Phased-Array Transceiver with On-Chip antennas in silicon: Receiver and Antennas," *IEEE J. Solid-State Circuits*, **41**(12), 2795-2806, Dec 2006.

[3] A Natarajan, A Komijani, X Guan, A Babakhani, and A. Hajimiri, "A 77-GHz Phased-Array Transceiver with On-Chip antennas in silicon: Transmitter and Local LO-Path Phase Shifting," *IEEE J. Solid-State Circuits*, **41**(12), 2807-2819, Dec 2006.

[4] H. Hashemi, X. Guan, A. Komijani, and A. Hajimiri, "A 24-GHz SiGe phased-array receiver - LO phase shifting approach," *IEEE Trans. Microwave Theory Tech*, **53**(2), 614-626, Feb 2005.

[5] A. Babakhani, D. B. Rutledge, and A. Hajimiri, "Transmitter Architectures Based on Near-Field Direct Antenna Modulation," *IEEE Journal of Solid-State Circuits*, vol. **43**(12), 2674-92, Dec 2008.

"This (practically) unlimited number of high-speed transistors combined with the crossing-over of chip-size and cut-off wavelength has produced a unique opportunity that is just beginning to be harnessed."

Wireless Networks

Communication and Control

by Babak Hassibi

Due to recent advances in wireless technology and RF circuitry, wireless devices (and other devices enhanced with wireless capabilities) have become more and more ubiquitous. By all appearances, this trend will continue. Furthermore, with advances in sensor technology, it is now possible to deploy a multitude of small sensors capable of exchanging information in a wireless fashion. With minimal infrastructure, bandwidth and power, such devices can provide mobile ad-hoc networks (MANETs) and sensor networks that deliver nearly arbitrary amounts of relevant information to any location at any time. The realization of such powerful sensor networks presents many research challenges since the networks need to self-organize, operate in a distributed fashion, and be robust with respect to changes in the network topology, time-varying traffic types, etc. The work supported by the Lee Center in the Hassibi group, as further explained below, has attempted to address some of these research challenges.

> "mobile ad-hoc networks (MANETS)...that deliver nearly arbitrary amounts of relevant information to any location at any time."

One normally thinks of a network as a collection of nodes with edges connecting them. However, in a wireless network, things are more complicated and we need to dispense with this picture. In particular, wireless networks have the following distinguishing features: shared communication medium; interference, i.e., the received signal on each node is the sum of all incoming signals; broadcast, i.e., the transmit signals on all outgoing edges are identical; path-loss, i.e., the signal strength depends on the distance between nodes so that both geometry and topology are important; and fading, i.e., the signal strength undergoes random fluctuations due to changes in the environment, such as might be caused by the mobility of sensors. Traditional methods have viewed these as "deficiencies." However, new approaches might be able to exploit them for improved performance.

The first wireless networks (deployed in the 1950's) consisted of a single base station with large coverage area in an urban center, along with mobiles. However, in these early systems only a few users could be accommodated due to interference. The theory of cellular networks was developed at Bell Labs in the 60's and 70's. This theory used path-loss to allow for frequency re-use, and thereby could accommodate many, many more users. However, it was not until the 1980's that the first such system was deployed by Motorola in Chicago (called AMPS—American Mobile Phone System). An interesting anecdote is that in the 1970's AT&T hired McKinsey & Company to study the feasibility of deploying a cellular system, but they were discouraged from doing so because McKinsey believed there was not a large enough market! Since the first commercial systems, cellular networks have rapidly evolved in terms of capacity, cell size, and access schemes. We now also have next generation wireless networks such as WLAN's (802.11's), blue tooth, Wi-Fi, Wi-Max, etc., along with the emerging sensor and ad hoc networks mentioned earlier.

Some Research Highlights

In what follows, we will briefly describe some of the research activities performed under the support of the Lee Center.

(1) Multiple antennas in cellular networks: To increase the capacity of cellular systems, it has been proposed to use multiple transmit antennas at the base station and (possibly) multiple receive antennas at the mobile users. We have studied the capacity of such a network and have shown that if each cell has a base station with M transmit antennas and there are n mobile users, each with N receive antennas, then the information-theoretic capacity scales as $C = M \log\log nN$. This shows that additional hardware investment at the base station is worthwhile, since the capacity increases linearly with the number of transmit antennas M, however, additional hardware expenditure at the mobiles is not, since the capacity increases only double-logarithmically in the number of receive antennas N.

A major challenge in delivering this capacity is that the transmitter needs to know the channel state information to all the n mobile users. This requires an unreasonable communication overhead and is not feasible in situations where due, for example, to mobility, the channels change rapidly. We have proposed a scheme (referred to as random beam-forming) which comes very close to achieving the information-theoretic capacity, yet requires minimal communication overhead. Random beam-forming, and its variants, are currently used in the Wi-Max standard. We have also studied scheduling problems in such systems (these become especially challenging, since there are not only variations in the users' demands, but also in the users' channel conditions which determine the rates to which they can be transmitted). In particular, we have studied the trade-off between the capacity of the system and the fairness in serving the different users and show that if the number of transmit antennas at the base station is on the order of the logarithm of the number of users, formally $M = O(\log n)$, then one can simultaneously both maximize the capacity and achieve fairness.

(2) Privacy in wireless networks: Since wireless is a shared medium, it is particularly susceptible to eavesdropping and therefore security is a major concern. An information-theoretic approach to guaranteeing security is through the notion of secrecy capacity, defined as the maximum rate at which two users can reliably communicate so that an eavesdropper can infer zero information from the communication of the two users. We have computed the secrecy capacity of a system in which the users, as well as the eavesdropper, have multiple antenna transmitters and receivers (as is the case in a cellular network). The solution of this long-standing, open problem has generated a great deal of follow-up research, and a large effort is underway to construct practical schemes that achieve this information-theoretic secrecy capacity.

(3) Capacity of wireless erasure networks: Unlike point-to-point information theory with a single transmitter and single receiver, where the information-theoretic capacity is well known and completely characterized, the information-theoretic capacity of almost all multi-user systems (with multiple transmitters, receivers and relays) is unknown. By fiat, the capacity of wireless networks in the general case is not known. We have recently proposed a simplified model of wireless networks, which we refer to as wireless erasure networks, where the transmitted packets are either perfectly received across links or dropped altogether with some probability. This applies in packet-based communication schemes (such as Internet protocols with in-packet error detection). With the model we have been able to compute the multi-cast capacity for transmission of the same information to multiple destinations in the network. The ideas used are similar to those used in network coding and the capacity is essentially determined by a certain "min-cut". This was arguably the first non-trivial class of wireless networks for which a capacity result could be obtained and it has generated significant further work in the research community.

(4) Estimation and control over lossy networks: Many complex systems (such as the power grid and many chemical plants) are controlled remotely with measurement and control signals being sent across a possibly lossy network. The random losses that occur in the network (both in terms of the measurements being dropped before being received by the estimator, or the control signals being dropped before reaching the actuators) have an obviously adverse effect on the system performance (and perhaps its stability). We have studied the design of optimal estimators and controllers in the face of such a lossy network. We have also determined universal laws (akin to the central limit theorem) that occur in such complex lossy control systems.

Future Work

The convergence of communications and signal processing, or rather communications and computation, has long been recognized. At an abstract level this is evident through the use of personal computers as communications devices (e.g., email and internet). Likewise, the relationship between signal processing and control has long been recognized. With recent technological advances, a complete convergence of communications, computation and control has begun.

Nowhere is this convergence more evident than in the mobile wireless networks that will emerge in the coming decades. As mentioned earlier, due to advances in sensor technology and radio-frequency circuitry, wireless devices will have the capabilities of sensing and exchanging information in a centralized and also distributed and ad-hoc way. Moreover, as these devices couple with more complex systems, they will be able to function as actuators and thereby influence the environment in which they operate. This opens up a host of exciting possibilities in environmental monitoring and control, surveillance, flight formation, as well as decentralized sensing and control of industrial plants, the power grid, and traffic. It also presents many research challenges that we will continue to pursue. Two major ones are now described.

where the devices need to make real-time decisions but communications takes place over unreliable, time-varying channels. LEE

Babak Hassibi is Professor and Executive Officer for Electrical Engineering.

"With recent technological advances, a complete convergence of communications, computation and control has begun. "

Network information theory: A stumbling block in the design of wired and, especially, wireless networks is the paucity of results in network information theory. In fact, the information-theoretic capacity of even the simplest networks remains open. Therefore much of the research in the community has focused on very specific models that are amenable to analysis (such as the wireless erasure networks mentioned above), or on asymptotic results that study various scaling laws for very large networks.

We have developed a new formulation of network information theory problems that reduces them to solving a linear optimization problem over the convex space of entropy vectors. The space of entropic vectors is a convex cone of 2^n-1 dimensions (where n is the number of random variables) and, if known, reduces network information theory problems to simple convex optimization. The challenge is that a characterization of this space is not known for n>3 and so our observation does not immediately solve network information theory problems; it rather reveals its core. Among others, this problem has connections with group theory, matroids, lattice theory, and determinantal inequalities. We have been actively studying this space and to date have constructed various inner and outer bounds.

Real-time information theory: Information theory, as developed by Shannon in the 40's and 50's and which serves as the theoretical core of today's telecommunications industry, ignores real-time constraints and allows for arbitrarily long delays in decoding the transmitted information. Control theory, on the other hand, ignores information-theoretic considerations and yet deals with real-time constraints upfront. We are therefore studying ways to combine communications and control theory in the context of a wireless network

Read more at:
http://www.systems.caltech.edu/EE/Faculty/babak

References

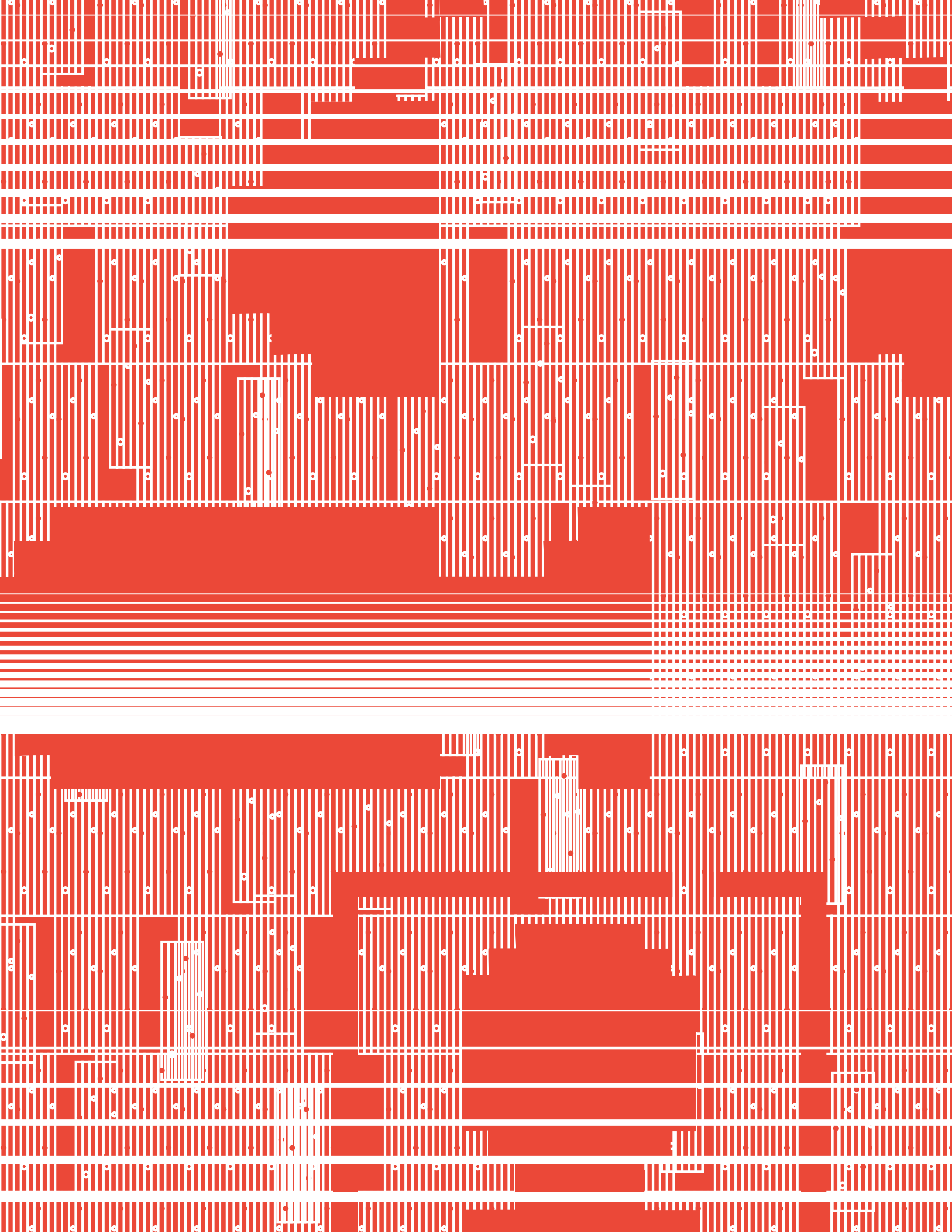
[1] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial channel state information," *IEEE Transactions on Information Theory*, Vol. **51**, No. 2, pages 506-522, (February 2005).

[2] F. Oggier and B. Hassibi, "On the secrecy capacity of the MIMO wiretap channel," *IEEE Transactions on Information Theory*, (2010).

[3] A.F. Dana, R. Gowaikar, R. Palanki, B. Hassibi and M. Effros, "Capacity of wireless erasure networks," *IEEE Transactions on Information Theory*, Vol. **52**, No. 3, pages 789-804, (March 2006).

[4] R. Gowaikar, B. Hochwald and B. Hassibi, "Communication over a wireless network with random connections," *IEEE Transactions on Information Theory*, Vol. **52**, No. 7, pages 2857-2871, (July 2006).

[5] V. Gupta, A.F. Dana, J.P. Hespanha, R.M. Murray and B. Hassibi, "Data transmission over networks for estimation and control," *IEEE Transactions on Automatic Control*, Vol. **54**, No. 8, pages 1807-1819, (2009).



Towards the Limits of Network Performance

Coding and Information Theoretic Approaches for Efficiency and Robustness

by Tracey Ho

Many features in modern networks have their roots in technologies developed for the Internet. The recent concept of network coding is an exception; and goes beyond traditional forwarding of immutable packets, viewing network operation instead as flows of information that can be operated on mathematically within the network. In particular, network nodes may perform mathematical operations across packets received on different incoming links so as to form new packets containing coded combinations of the original packets' information, all of which will be subsequently decoded at other nodes in the network. Traditional store-and-forward operation, where network nodes simply route packets unchanged from incoming to outgoing links, is a special case of this more general network coding framework.

"...fundamentally improve network performance, particularly in uncertain, adversarial, or resource-constrained scenarios. "

Early work on network coding by myself and others had demonstrated promising aspects of network coding under a small collection of network scenarios. Recent work from my group at Caltech, supported by the Lee Center, has broadened the range of network coding techniques as well as the range of applicable network types and scenarios. Our research investigates the use of network coding and other information theoretic approaches so as to fundamentally improve network performance, particularly in uncertain, adversarial, or resource-constrained scenarios.

One application of network coding is in network error correction, where redundant information is coded across multiple network links. This enables reliable communication over a network even in cases where individual links cannot be made reliable; for instance, in case of adversarial attacks or arbitrary faults, for which it is not sufficient to do error correction on a link by link basis. Prior work had considered network error correction for the case of single-source networks that are homogenous in the sense that each link (or packet) has equal capacity and is equally vulnerable to errors. We broadened the scope to include multiple-source networks and heterogeneous links, and thereby showed that in-network error correction coding allows redundant capacity to be shared among multiple sources[1]. Notably, such coding makes it possible to achieve the same transmission rates as if each source were given exclusive use of all the redundant capacity. This is illustrated in Figure 1, where if we allocate part of the network to carry information from source 1 only, and the remainder of the network carries information from source 2 only, then either one of the sources, but not both, is able to communicate reliably at rate 1 under a single link error. Coding at the intermediate nodes allows the two sources to use shared network capacity to send redundant information, so that both sources can simultaneously communicate reliably at rate 1 under a single link error. We also showed that non-homogenous networks require a wider range of different error correction coding strategies compared to the homogenous case. In a homogenous network, the error correction capacity can be achieved by linear network coding, and for a single source and sink, it is sufficient to do coding at the source and simple forwarding at intermediate nodes. However, in the non-homogenous case, we showed that linear coding is suboptimal in general, and that forwarding at intermediate nodes is not sufficient even for a single source and sink—instead, depending on the topology, coding, partial error detection or partial error correction at intermediate nodes could be required.

Besides security against errors, coding in networks is also useful for information security against eavesdropping, where coding is used to ensure that an eavesdropper observing a subset of network links receives no information about the message being communicated. For this problem, we also showed[2] that non-homogenous networks require a wider range of different coding strategies compared to homogenous networks, and that it is more complex to find the capacity limits and optimal strategy in the general case.

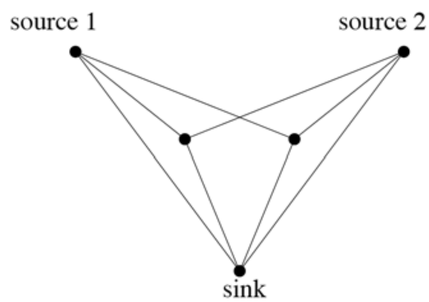


Figure 1: Example showing that in-network error correction coding is needed to share redundant capacity across multiple sources.

One limitation of purely information theoretic approaches is that they do not take advantage of computational limitations of adversaries. Cryptographic techniques exploit the assumption of computationally bounded adversaries, but at the same time place a higher computational burden on the communicating nodes, which becomes an important limiting factor on performance in networks with computationally limited nodes, such as wireless and sensor networks. Our recent research is merging information theoretic coding and cryptographic techniques for robustness against a wide range of natural impairments and adversarial attacks. We have developed hybrid strategies, which take advantage of the assumption of computationally bounded adversaries, to provide secure and efficient communication when some or all of the communicating nodes are computationally limited. In such strategies, different packets may undergo different operations at a node, and different nodes may employ different techniques based on their topological location and capabilities. For instance, we developed an adaptive network error correction strategy which, when used in conjunction with probabilistic cryptographic verification of subsets of packets processed by network nodes, can achieve a higher rate of reliable information compared to purely information theoretic or purely cryptographic approaches[3]. Our eventual goal is to provide practical guidance on how to combine various information theoretic coding techniques and cryptographic techniques optimally under different network scenarios and constraints such as limitations in power and processing speed of different network nodes; and to develop efficient algorithms that adapt dynamically in the presence of uncertain and changing network conditions.

Another application of network coding is in distributed network storage. By distributing important information across different storage nodes with sufficient redundancy, information retrieval can be made robust to node failures, loss of connectivity, data corruption and so on. There is a fundamental trade-off between the dissemination/storage costs and the reliability of information retrieval. Similar to the robust transmission problem, the cost-reliability trade-off in the robust storage problem can be improved by use of coding. While exact optimization of the cost-reliability trade-off, given an arbitrary network topology and failure/mobility model, is very complex, we have obtained exact solutions for some special cases (e.g. the case where storage costs dominate and the storage budget is large enough to support a high probability of successful information retrieval from a random subset of storage nodes of given size[4]) and approximate approaches for other cases. We are also investigating regeneration of stored information following failures or errors among the storage nodes. We have found new constructions of practical systematic codes that can be regenerated with minimum data downloaded from other storage nodes, and our ongoing work seeks to obtain a characterization of the set of all possible systematic codes for a given number of storage nodes, code rate and repair bandwidth, as well as an efficient method to systematically enumerate these codes.

"merging information theoretic coding and cryptographic techniques"

My group is also studying the design of new practical wireless networking techniques using network coding and information theoretic approaches. For example, two-way relaying scenarios where two or more terminal nodes exchange information via a common relay node provide a canonical instance of the usefulness of network coding, and can be used as building blocks for practical code construction in general networks. We showed how to generalize any given network layer coding strategy to a family of joint physical-network layer strategies where the relay effectively propagates physical-layer soft information to improve decoding at the terminals, and also showed how the optimal strategy is affected by the channel parameters and relative power constraints of the relay and terminals[5]. Another result we obtained was a new class of random medium access control protocol, which allows each user to transmit at multiple data rates and uses successive interference cancellation so that multiple packets can be received simultaneously. We found that the proposed protocol can achieve a significant throughput gain over existing random medium access protocols under a Gaussian wireless model. We plan to consider the case of fading wireless channels as the next step towards realizing the benefits of this new approach in practice. L E E

Tracey Ho is Assistant Professor of Electrical Engineering and Computer Science.

Read more at: http://www.its.caltech.edu/~tho

References

[1] S. Vyetrenko, T. Ho, M. Effros, J. Kliewer, and E. Erez, "Rate regions for coherent and noncoherent multi-source network error correction," in *IEEE ISIT*, June 2009.

[2] T. Cui, T. Ho, and J. Kliewer, "On secure network coding over networks with unequal link capacities," *IEEE Transactions on Information Theory*, 2009, submitted.

[3] S. Vyetrenko, A. Khosla, and T. Ho, "On combining information theoretic and cryptographic approaches to network coding security against the pollution attack," in Asilomar Conference on Signals, Systems, and Computers, invited paper, November 2009.

[4] D. Leong, A. Dimakis, and T. Ho, "Distributed storage allocation for high reliability," in *IEEE International Conference on Communications*, 2009, to appear.

[5] T. Cui, T. Ho, and J. Kliewer, "Memoryless relay strategies for two-way relay channels," *IEEE Transactions on communications*, **57** (10). pp. 3132-3143, October 2009.

Networking Lab

Understanding Behavior and Control of the Internet

by Steven H. Low

The Internet has become, in a short span of four decades, critical infrastructure for the world and a common platform for innovation with impact far beyond communications. It connects 20% of humanity and offers a trillion websites as of 2008. Underlying the apparent convenience and simplicity, however, is a complex array of hardware, software, and control algorithms. Indeed the Internet is the largest distributed nonlinear feedback control system ever built. Despite the Internet's importance, there is no fundamental understanding of its global behavior and future evolution. Will congestion-induced collapse, first detected on the Internet in 1986 with only 5,000 hosts, reoccur in the modern network with more than half a billion hosts? How can we optimize the interaction between congestion control, routing, and the wireless infrastructure? Will performance and stability problems arise as the Internet continues to scale in size, speed, scope, heterogeneity and complexity?

Although the working elements (hardware, software, and algorithms) that make up the Internet are well understood, their collective behavior is much harder to quantify, and is one of the defining features of a complex system. A comment made by Caltech's Gordon and Betty Moore Professor, Emeritus, Carver Mead, some 20 years ago about computational neural systems applies equally well to the Internet today: "The complexity of a computational system derives not from the complexity of its component parts, but rather from the multitude of ways in which a large collection of these components can interact. Even if we understand in elaborate details the operation of every nerve channel and every synapse, we will not by so doing have understood the neural computation."

Developing a fundamental understanding of large-scale networks remains a most critical and difficult challenge in networking research. It is this interconnection that provides the myriad of communication services that we now take for granted, but that also allows the failure of a single power plant to black out a continent or the crash of a stock to trigger a regional financial meltdown. A comprehensive theory of large-scale networks is the only certain way to harvest their power and manage their risk.

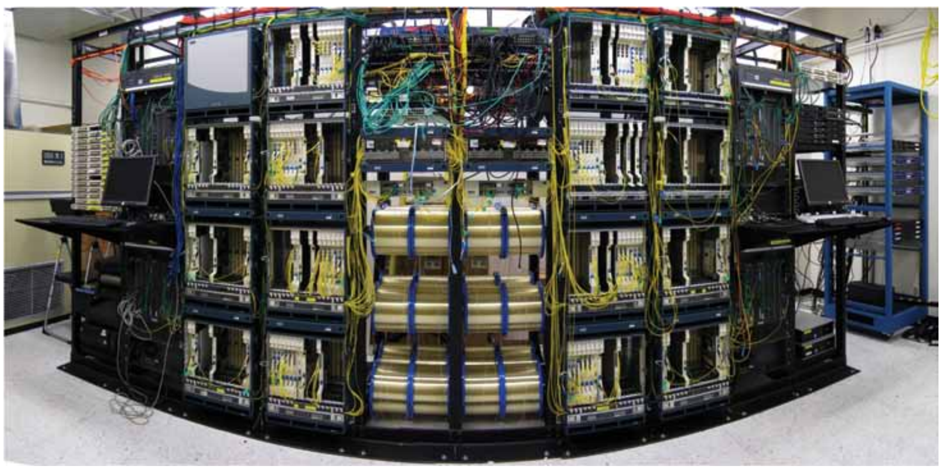


Figure 1: Caltech WAN-in-Lab testbed (2008).

Since 2000, Caltech's Networking Lab has established an integrated research program where theory, experiment, and networking infrastructure can intimately inform and influence each other. We have been developing a mathematical theory of the Internet, focusing particularly on congestion control. One outcome has been the invention of a new congestion control algorithm called FAST TCP. Moreover, we have built a unique testbed (WAN-in-Lab, see Figure 1) to test our theory, and have also started a company FastSoft to drive the deployment of FAST TCP. Our research program has had two main results. First, it has firmly established a theoretical basis for TCP (Transmission Control Protocol) design, and has brought mathematical rigor to TCP research. Second, it has demonstrated the feasibility of scaling TCP to very high speed global networks of the future and has developed some mathematical tools and design approaches to develop and evaluate new protocols.

Theory

With Professor John Doyle, we have developed a theory of Internet congestion control based on convex optimization and control theory[1, 2, 5]. In this theory, congestion control is a distributed asynchronous primal-dual algorithm carried out over the Internet in real-time to solve an abstract global optimization problem. Different TCP (Transmission Control Protocol) algorithms differ merely in the objective functions they implicitly optimize. This theory allows us to understand the limitations of the current TCP and to design new algorithms. Until recently, the majority of TCP research has largely been simulation-based, often using only a single bottleneck link and a single class of algorithms. Our theory can predict the equilibrium behavior of an arbitrary network operating under any TCP-like algorithms. Moreover, for the first time, we can predict and design efficiency and stability properties in the presence of feedback delay of arbitrary networks.

The theory has also led to the discovery of some counterintuitive behaviors of large-scale networks. For instance, a network with heterogeneous congestion control algorithms can have multiple equilibrium points, and not all of them can be locally stable. Such phenomena will become more important as the Internet becomes more heterogeneous. It is generally believed that a resource allocation policy can either be fair or efficient but not both. We characterize exactly the tradeoff between fairness and throughput in general networks. The characterization allows us both to produce the first counter-example and trivially explain all the previous supporting examples in the literature. Surprisingly, our counter-example has the property that a fairer allocation is always more efficient. Intuitively, we might expect that increasing link capacities always raises aggregate throughput. We show that not only can throughput be reduced when some link increases its capacity, but it can also be reduced when all links increase their capacities by the same amount. Often, interesting and counterintuitive behaviors arise only in a network setting where flows interact through shared links in intricate and surprising ways. Such behaviors are absent in single-link models that were prevalent in the TCP literature; and they are usually hard to discover or explain without a fundamental understanding of the underlying structure. Given the scale and diversity of the Internet, it is conceivable that such behaviors are more common than we realize, but remain difficult to measure due to the complexity of the infrastructure and our inability to monitor it closely. A mathematical framework thus becomes indispensable in exploring structures, clarifying ideas, and suggesting directions. We have demonstrated experimentally some of these phenomena using the WAN-in-Lab testbed.

Experiment

We have implemented the insights from theoretical work in a new protocol FAST TCP[4] and have worked with our collaborators to test it in various production networks around the world[3]. Physicists, led by Professor Harvey Newman of Caltech's Division of Physics, Mathematics and Astrnomy, have been using FAST TCP to break world records on data transfer from 2002–2006. Perhaps more important than the speeds and data volumes of these records, has been the impact of our experiments on the research and application of TCP. For instance, the Internet2's Land Speed Record in late 2002 was attained in an experiment where data was transferred for only 16 seconds—the techniques that were used in that, and all previous, records, did not address protocol issues and were therefore so fragile that high throughput could be sustained only momentarily. As soon as network fluctuations occurred, throughput would collapse, and the experiment was terminated. Our experiments in the 2002 SuperComputing Conference were the first to sustain line rate (at mutlti-Gbps) for several hours, using a standard packet format. At that time, the high performance computing community was very skeptical about the feasibility of using TCP as the transport protocol for high-speed (multi-Gbps) long-distance (4,000 miles) transfer of large volumes (petabytes, or 10^{15} bytes) of scientific data. Our experiments settled the debate and changed the direction in the high-energy physics community from developing

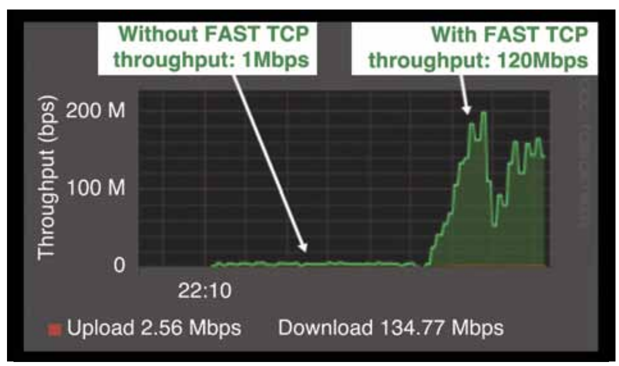


Figure 2: In June 2007, part of Sprint network suddenly suffered up to 15% packet loss for about 15 minutes. During this period, the throughput between FastSoft's San Francisco data center and New York data center dropped to 1 Mbps. FastSoft acceleration provided extreme resilience to packet loss and restored the throughput to 120 Mbps.

UDP-based data transfer solutions to TCP-based solutions. This and subsequent experiments have been widely reported in the media including *Nature, National Geographic, New Scientist, The Economist, BBC, CNN, Reuters, Business Week, The Australian, MSNBC, Yahoo! News, PC Magazine, ComputerWorld, CNet,* etc.

Testbed

We have built a one-of-a-kind academic testbed, WAN-in-Lab, at Caltech for the design, development, and evaluation of high speed network protocols (see Figure 1). It uses real carrier-class networking hardware to avoid the artifacts introduced by network simulation and emulation, while being localized to allow detailed measurement of network behavior. WAN-in-Lab, funded by the Lee Center, NSF, ARO, Cisco and Corning, is literally a wide-area-network—with 2,400 km of long-haul fibers, MEMs optical switches, wavelength division multiplexing equipment such as optical amplifiers and dispersion compensation modules, routers, servers and accelerators. It provides a unique platform where new algorithms can be developed, debugged, and tested first on WAN-in-Lab before they are deployed in the field, thus shortening the cycle of design, development, testing and deployment.

Deployment

We have spun off a company, FastSoft, in 2006 to drive the development and deployment of FAST TCP. FAST TCP forms the core of a set of web acceleration technologies developed at FastSoft for the delivery of web applications and contents over the Internet to enterprises and consumers. Such technologies will become increasingly important with the proliferation of video, dynamic content, and cloud services. They are accelerating the world's second largest content distribution network and other Fortune 100 companies, typically by 1.5 to 30 times. LEE

Steven H. Low is Professor of Computer Science and Electrical Engineering.

Read more at: http://netlab.caltech.edu

References

[1] S. H. Low, "A Duality Model of TCP and Queue Management Algorithms," *IEEE/ACM Trans. On Networking*, **11**(4):525-536, August 2003.

[2] F. Paganini, Z. Wang, J. C. Doyle, and S. H. Low, "Congestion Control for High Performance, Stability and Fairness in General Networks," *IEEE/ACM Trans. on Networking,* **13**(1):43–56, February 2005.

[3] C. Jin, D. X. Wei, S. H. Low, G. Buhrmaster, J. Bunn, D. H. Choe, R. L. A. Cottrell, J. C. Doyle, W. Feng, O. Martin, H. Newman, F. Paganini, S. Ravot, and S. Singh, "FAST TCP: From Theory to Experiments," *IEEE Network*, **19**(1):4-11, January/February 2005.

[4] D. X. Wei, C. Jin, S. H. Low, and S. Hegde, "FAST TCP: Motivation, Architecture, Algorithms, Performance," *IEEE/ACM Transactions on Networking*, **14**(6):1246-1259, December 2006.

[5] Mung Chiang, Steven H. Low, A. Robert Calderbank, and John C. Doyle, "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures," *Proceedings of the IEEE*, **95**, 255-312, January 2007.

Market Research in the Lee Center

by Charles R. Plott

Under Lee Center support, we have conducted three related research projects relating to markets. The first contribution is a new function for market-like institutions that exists on the perimeter of information and complex systems. Its success is supporting the development of new businesses. The second is basic science, illustrating the close relationship that exists between economics and network systems. It is a step towards solving one of the oldest mysteries in economics. The third is a contribution to methodology and procedures, demonstrating how to create and study a phenomenon that has been elusive in both laboratory and field studies.

Information Aggregation Mechanisms

Information aggregation mechanisms are "market like" processes that collect "soft" information held by dispersed individuals in the form of intuition and hunches and then transform it to quantitative form. The basic principles of behavior can be seen working daily in the stock market as well as other markets that have widely recognized capacities to predict events. Terms like "the market thinks that X" reflect the fact that underlying economic events known or suspected by only a few, such as the health of a business, the outcome of a pending law suit, company earnings, etc., are frequently forecast by market activity. Research funded by the Lee Center contributed to the identification and characterization of these principles and their application to the design of new types of mechanisms that have only a single purpose, the gathering of targeted information that is otherwise widely dispersed in the form of hunches and intuition by many individuals.

This research has resulted in a substantial characterization of the principles at work in the information aggregation mechanism. The characterization has lead to theories of the types of processes that might be successful, the circumstances under which the mechanism might be successful, the measurement of the reliability of the information produced by the mechanism and procedures for managing the application of the mechanism so as not to diminish performance.

Examples of applications include forecasting: future product sales based on the opinions of sales force, the revenue of upcoming movies, the price of plywood, economic indicators, and even the proposals for the U.S. Treasury bailout auction. The research supported by the Lee Center has evolved into a broad range of activities that involves other researchers and other fields. A journal and professional meetings have emerged specializing in the area. A popular book, *The Wisdom of Crowds*, has also drawn attention to the area. Businesses that were specialized in closely related activities have altered their business plans to more closely mirror findings of the scientific community. Examples are Intrade Prediction Markets and the Hollywood Stock Exchange.

The Dynamics of Price Discovery in Markets

A second research project was focused on the dynamics of market price discovery. For decades the science has been challenged by a fundamental problem. How is it that prices evolve through competition in a market?

Typically, market systems are modeled as a large set of simultaneous equations that no one knows. From this view three questions emerge. Can markets solve the system of equations? When it happens, how does it happen? Is the phenomenon robust? Technology developed with the support of the Lee Center facilitated experiments with multiple, interacting markets and demonstrated that "numerous" and "interdependent" markets tend to converge to the equilibria of the competitive model. The most fundamental discovery is that within certain market organizations, the dynamics of the price discovery process for converging to the stable equilibria, appears to be similar to the Newton method of finding solutions to systems of equations. Information about derivatives needed for the Newton method enter through the order book.

The demonstration that the mystery of market price discovery resides in the dynamics of the price formation process immediately calls forth another question. Do these basic principles continue to apply when parameters that influence buyers and sellers change randomly? Experiments were conducted in which markets were subject to continuous shocks, as if traders randomly appear, do their business and disappear. The results are that two different, but related laws of supply and demand operate. First, there is a temporal equilibrium, resulting from the classical law of supply and demand applied to traders who are in the market at any instant. Prices will be higher if a large number of buyers together with a small number of sellers happened to show up. This temporal equilibrium, which reflects the buyers and sellers who are in the market at the moment, is biased toward a flow equilibrium (a price that balances the expected demand and supply given the stochastic structure of parameter changes). The flow equilibrium can be viewed as a type of "long run" equilibrium even though prices never stabilize at the flow equilibrium level. The two laws interact such that "on average" prices equal the long term equilibrium but at any instant prices are closely related to the temporal equilibrium, with a bias away from the temporal equilibrium in the direction of the long term.

Tacit Collusion and Collusion Breaking Institutions

A concept of "tacit collusion" preoccupies economics, the legal (anti- trust) profession and the courts. It is a type of breakdown in competition that economic theory seeks to define and characterize and for which the law seeks legal remedies. The theory is that competitors can "tacitly" collude by simply recognizing their common interest and acting accordingly to restrain competition, as if they had actually participated in a conspiracy. The outcome of tacit collusion will not be the competitive, Nash equilibrium but will instead be a collusive equilibrium in which competitors will be better off and consumers will be worse off.

From a scientific point of view it is a rather mysterious phenomenon. Prior to the Lee Center project, it had not been observed directly and had been measured/detected only in terms of deviations from what would be expected if tacit collusion did not exist. Indeed, tacit collusion had never been convincingly observed in experiments.

The Lee Center project stems from the insight that complexity and interdependencies of markets might be a contributing factor to tacit collusion since coordination and specialization of competitive functions, key features of tacit collusion, evolve naturally as part of competitive interactions. The study called for the creation of an environment in which tacit collusion would readily emerge by putting together all factors thought to contribute. Then, if tacit collusion emerged and began operating, institutional changes that would make the collusion dissolve would be introduced—an example of "remedies" as sought by the courts.

The effort to create tacit collusions in the laboratory was extremely successful, with tacit collusions emerging readily in the context of multiple, simultaneous, ascending price auctions in which different items are sold to the competitors. The interdependence of the system allows the targeting of punishment strategies that participants aggressively and selectively employ to shape individual competitors into collusive conformity. The set of conditions is termed "the collusion incubator environment" and has been used repeatedly in the experimental literature to study the process of collusion formation and the conditions that might produce collusion.

The research has also considered ways to stop collusion and return the system to the competitive equilibrium. It was found that almost all remedies suggested by popular theory and the law fail to restore the system to a "competitive" mode of behavior. Guided by subtle aspects of theory, attention then moved to the implementation of a structural change from simultaneous increasing price auctions to simultaneous descending price auctions. The decreasing price auction (first to take as price goes down) removes the ability for competitors to "punish" (or even react to) those who do not conform to collusive practices. This institutional change immediately destroys the collusions. Current analysis is focused on the processes and mechanisms through which the collusion is controlled in the hope that new tools and insights are provided to businesses that must be sensitive to the possibility of collusion within those with whom they deal; as well as the government that is responsible for protecting the public by maintaining competition. LEE

Charles R. Plott is Edward S. Harkness Professor of Economics and Political Science.

Radio Transmitters

Amplifiers for Wireless Networks

by David B. Rutledge

In network jargon, the transmitter is part of the physical layer. It sends power to the other nodes through an antenna. In many wireless systems, transmitters are the most expensive components. One thing that increases the cost in cellular data and voice transmitters is that the power amplifiers are often made of gallium arsenide rather than silicon. Another is that traditional computer aided design software has not been adequate for transmitters. Instabilities are often discovered late in the development cycle when they are difficult to correct. In many wireless systems, the transmitter consumes most of the power and produces most of the heat. Yet another problem is that, historically, transistor-based transmitters have not produced enough power for high-speed satellite network uplinks. Our research has addressed many of these problems. Much of the work is in collaboration with Ali Hajimiri, Professor of Electrical Engineering at Caltech and Lee Center Visitor Professor Almudena Suarez.

In 2001, two of my students, Scott Kee and Ichiro Aoki, developed a new efficient class of switching amplifier called E/F (Figure 1). This work was followed by a series of papers with Scott Kee, Ichiro Aoki, and Ali Hajimiri that incorporated this class of switching amplifier into a power combining circuit that used a transformer with transistor pairs at the corners. This increased the power and efficiency to such an extent that we were able to demonstrate a silicon integrated circuit that had high enough output to act as power amplifier for a cellular telephone. Ali, Scott, and Ichiro founded a company called Axiom Microdevices to develop this power amplifier for the GSM cellular phone market. The company was eventually acquired by Skyworks, and this amplifier is one of their big sellers, with cumulative sales of over 50 million units.

Almudena Suarez from the University of Cantabria visited Caltech on sabbatical in 2004. Almudena is an expert in the development of computer software for non-linear microwave circuits. She had worked extensively with low-power mixers and oscillators, but until that time had not worked with high-power transmitters. Two of my students, Sanggeun Jeon and Feiyu Wang, had been developing high-power transmitter circuits for radar and industrial applications. Sanggeun and Feiyu had found that the circuits showed many instabilities that were not predicted by the computer-aided design software available at that time. There were oscillations away from the drive frequency, jumps in the output power, and increased noise in some drive power ranges. These instabilities have been known in industry for almost a hundred years. When the instabilities show up, they are a disaster, because solutions must be discovered by trial and error. Sanggeun and Feiyu had filled an entire shelf with these strange power amplifiers. I called it "the zoo." When Almudena arrived, the three of them set to work in analyzing and trying to understand each of the instabilities. Eventually they succeeded in tracking every one down, and developed a way to detect each instability in the design software before the transmitter was built. This led to a series of three papers in the IEEE microwave journals[2,3,4] that laid out a design approach to avoid these instabilities. For this and other work, Almudena Suarez has become extremely influential in the microwave community, and I believe that these techniques will see wide use as people become more familiar with them.

Figure 1: Prototype E/F switching amplifier developed by Scott Kee and Ichiro Aoki[1]. This amplifier produced 1.1 kW at 7 MHz with 85% efficiency.

Frequencies above 15GHz are commonly used for mobile satellite uplinks because the antennas can be smaller than at lower frequencies. However, there is a serious challenge in using these frequencies. Specifically, to obtain high data rates, one needs more power than has traditionally been available from integrated circuits, typically 3Watts at these frequencies. While power levels in the range of 50Watts are available from a vacuum tube transmitter, making the data transmission faster, and giving better transmission through rain, vacuum transmitters are heavy and expensive. A vacuum-tube transmitter that we purchased for measurements in my lab cost $100,000 twenty years ago, and weighed over one hundred pounds. In addition, the vacuum-tube circuits often have burn-out problems. All of these difficulties have encouraged engineers to find an alternative to vacuum-tube transmitters in mobile satellite uplinks. What was needed was a way to combine the output power of a large number of transistors within an integrated circuit. Previous integrated circuits used transmission-line combiners that typically only allow the power from about eight devices to be combined efficiently. These transmission lines consume an increasing fraction of the power as the number of devices is increased, and that puts a stop to it. Students in my group invented a way to combine the outputs of several hundred or more transistors efficiently by letting

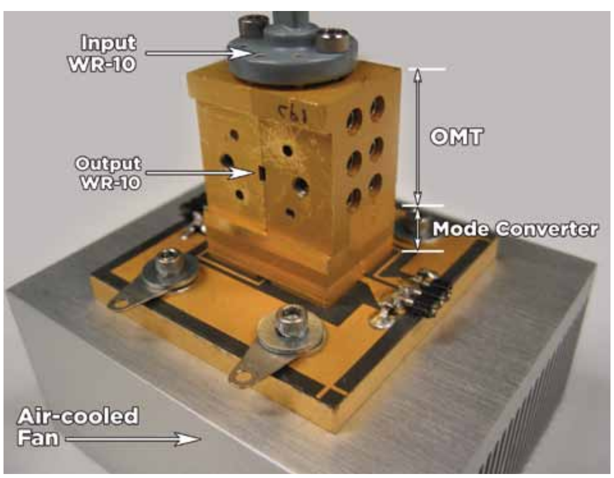


Figure 2: Grid amplifier operating at 79 GHz with an output power of 260 mW[5].

them radiate collectively into a metal waveguide. The device is called a grid amplifier. IST post-doctoral fellow Younkyu Chung, and Caltech students Chun-Tung (Lawrence) Cheung and Michael DeLisio showed that the approach could work at 79GHz, a very high frequency[5].

Lawrence, Mike, and another student from my group, Chad Deckman, started a company called Wavestream (see discussion of Wavestream later in this booklet) to develop and sell these grid amplifiers. The company is located in San Dimas and has 120 employees. Their best selling product is a 50-W, 30-GHz transmitter. Ten years ago, a transistor-based transmitter at this power level and frequency would have been inconceivable. LEE

David B. Rutledge is the Kiyo and Eiko Tomiyasu Professor of Electrical Engineering.

Read more at:
http://www.its.caltech.edu/~mmic/group.html

References

[1] Scot Kee, Ichiro Aoki, David Rutledge, "7-MHz, 1.1-kW Demonstration of the New E/F2, odd Switching Amplifier Class,'' *International Microwave Symposium, Phoenix*, May 2001.

[2] Sanggeun Jeon, Almudena Suarez, and David Rutledge, "Global Stability Analysis and Stabilization of a Class-E/F Amplifier," *IEEE Transactions on Microwave Theory and Techniques*, MTT-**53**, 3712–3722 (2005).

[3] Sanggeun Jeon, Almudena Suarez, and David Rutledge, "Analysis and Elimination of Hysteresis and Noisy Precursors in Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, **54**, 1096–1106 (2005).

[4] Feiyu Wang, Almudena Suarez, and David Rutledge, "Bifurcation Analysis of Stabilization Circuits in an L-band LDMOS 60-W Power Amplifier," *IEEE Microwave and Wireless Component Letters*, **15**, 12–714 (2005).

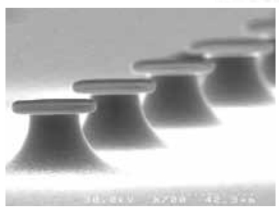
[5] Younkyu Chung, Chun-Tung Cheung, Michael P. DeLisio, and David Rutledge, "W-band Waveguide-Packaged InP HEMT Reflection Grid Amplifier," I*EEE Microwave and Wireless Component Letters*, **16**, 324–326 (2006).

High-Q for Photons

Nonlinear Resonators for Light

by Kerry J. Vahala

Since the mid 90s, information traveling as a lightwave on optical fiber has been multiplexed spectrally. One can imagine this process as information streams or channels traveling at distinct colors, with each color providing a channel of information—much like in television or radio transmission. Some time ago, the requirement to combine and separate these channels triggered interest in compact optical resonators[1]. Like their counterparts in electronics, these devices could be used to select or combine channels by the process of filtering signals according to frequency. Resonators, whether electronic, optical, or even acoustic, have in common the idea of energy storage at a particular frequency. The pitch of a tuning fork is a good example in the acoustic realm; and the purity of the pitch is a mathematical property that can be related to how long energy is stored. This storage time is conveniently given in terms of the number of cycles of vibration of the signal, and is called the "Q factor." A tuning-fork resonator, for example, will have a Q factor in the range of 10,000. Another important resonator, the quartz crystal used in electronics, can have Q factors as high as 1 million. Such a high purity-of-pitch makes quartz resonators useful in many applications, such as the familiar quartz wristwatch.



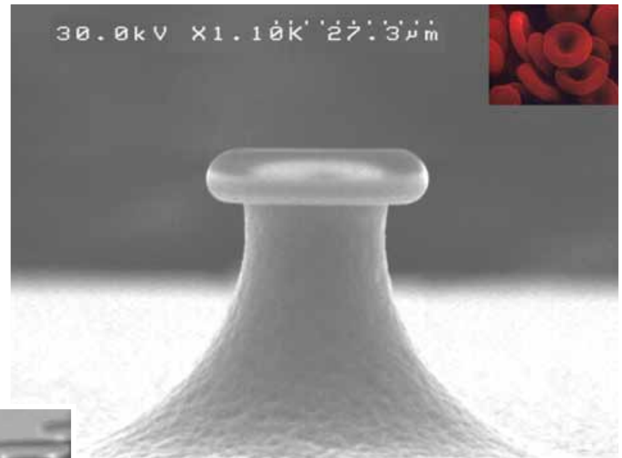


Figure 1: An array of silica microtoroids (lower left) and a close-up side-view of an individual device (main panel). Light orbits the perimeter of the donut-shaped toroidal structure and is confined for a relatively long time because of the smooth glass surface. For a size comparison, red blood cells are shown in the upper right panel.

In optical communications, the Q factor of the best semiconductor, chip-based resonators hovered around 10,000 until about 7 years ago. At that time, a technique for fabrication of optical resonators called microtoroids was developed my research group[2] . The toroid is fabricated of silica glass on silicon wafers that are similar to those used in electronics; and they hold light within a tiny ring-shaped orbit (see Figure 1). Normally, photons, the unit particle of light, are knocked from such orbits by imperfections in the glass such as the surface roughness created by the semiconductor processing, itself. However, through a special processing technique, the surface of these microtoroid resonators is smoothed to the atomic scale. The photons, in turn, enjoy a very smooth ride and remain in orbit a long time (high Q). Microtoroid resonators hold a record, even today, for on-chip devices. They feature Q factors in excess of 100 million. Moreover, because their fabrication relies upon many tools developed by the semiconductor processing industry, these Q values are obtained with a high yield so that many devices can be fabricated on a single chip.

> **"This leap in optical Q factor was so large that entirely new devices have become accessible using microtoroid resonators on silicon chips."**

This leap in optical Q factor was so large that entirely new devices have become accessible using microtoroid resonators on silicon chips. Narrow-band filters ("high purity pitch") with footprints smaller than the width-of-a-hair provide one example. However, we have also found that when energy storage can be boosted and concentrated into small volumes other effects can also occur. To understand some of these, imagine the photons of light as cars and their orbits within the microtoroid as a roundabout for the cars. The storage time (Q factor) of the microtoroid would determine how many times the cars must circle before leaving the roundabout. Clearly, as storage time increases, even a trickle of input traffic creates congestion. Such traffic congestion is made far worse in the microtoroid, by its small size. In fact, a trickle of input photons (meaning very-low input power) is sufficient to create such intense congestion that photon collisions become very likely. Unlike the collisions of automobiles, photon collisions can be precisely engineered to have useful outcomes (through control of the shape and size of the microtoroid). For example, in a process called third-order upconversion, three photons at a given wavelength can fuse together upon collision into a single, much higher energy photon. Using this process, microtoroids can generate visible light that is color tunable (see Figure 2) beginning with input infrared photons provided by compact laser sources made for the telecom industry[3].

In yet another kind of interaction, two photons can bump into one another; and exchange energy. In this fender-bender-like collision the photons emerge with new colors (one shifted to the red and other to the blue end of the spectrum): a fundamental process that enables a very useful device called a parametric oscillator. Through a combination of their high-Q, small volume, and shape-engineered properties, microtoroids achieved the first demonstration of these devices in any microcavity[4].

“Like the force a driver experiences when making a sharp turn, photons also apply a force on the glass interface of the microtoroid.”

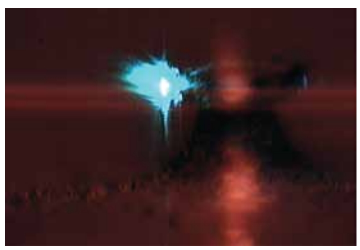


Figure 2: A nonlinear optical process called third-order upconversion is caused by the collision of three photons and the subsequent emergence of a single, higher-energy photon. In the upper panel, a single infrared beam of light is coupled into a microtoroid (out of focus in the panel) and blue (higher energy photons) are created. In the lower panel, two different wavelengths of infrared light are coupled in opposite directions into a single microtoroid (cross-section is sketched as dashed lines) and these produce two distinct, visible colors upon upconversion. The optical power level required for this process to occur is very low on account of the high storage time (high Q factor) of the microtoroid device.

Like the force a driver experiences when making a sharp turn, photons also apply a force on the glass interface of the microtoroid. Individual photons provide a negligible force, but here again because of the high density of photons, their collective force is sufficient to inflate the microtoroid by a small amount (a few picometers). Like a guitar string whose length is changed while being struck, this change in the toroid size also shifts the pitch (color) of the circulating light. Because the pitch depends on how many photons are circulating inside the microtoroid, two strange effects are possible. In a first, the photon pressure transfers energy to the mechanical structure and causes it to vibrate from radio to the microwave frequencies. As such, this first effect produces what is called regenerative or self-sustained oscillation, a very useful phenomenon at the heart of all communication systems. The fact that this oscillation is both optically powered, and the output signal is itself also a lightwave (albeit modulated with the oscillations) endows the device with insensitivity to normal electromagnetic interference effects[5].

The second effect emerges by slightly adjusting the wavelength of the applied photons. When this is done, the flow of energy can be reversed so that power flows from the mechanical structure of the microtoroid to the photons. This transfer of energy cools the mechanical motion and is now the subject of an intense, world-wide effort to achieve what is called the quantum ground-state of motion for an oscillator.

Looking to the future, one of the frontiers of communications is the question of how to best utilize the quantum properties of light for information transmission. In this context, the subject of cavity quantum electrodynamics provides a tool for reversible transfer of information stored on a photon into matter (atoms). A prerequisite for this transference is that optical losses must be greatly reduced so that the underlying quantum principles are dominant. In this regard, the microtoroid resonator provides a new approach to these studies. In collaboration with Professor Jeff Kimble in the Caltech Physics department, the Vahala group has helped demonstrate the first single-atom cavity QED measurement on a chip. This work offers a pathway to higher levels of functionality within the subject of quantum information studies. Besides these studies there are efforts underway to increase the functionality of ultra-high-Q devices including integrated waveguides as opposed to the optical fiber device now in use. Such an addition would greatly extend the range of applications of these remarkable devices.

Continuing Impact of Lee Center

Finally, if photons can travel in closed, circular paths for long periods of time (and hence effectively over a great distance) can they also be made to travel along open-ended, paths over the same distance on a semiconductor chip? The question is important as several applications would benefit. These range from bio sensors to high-performance synthetics filters and phased-array radars. The microtoroid result has challenged conventional thinking on optical loss for photonic waveguides; and created interest in "fiber-like-losses" for microphotonic circuits. These ideas are now being developed further at Caltech and elsewhere under support from DARPA. L E E

Kerry J. Vahala is the Ted and Ginger Jenkins Professor of Information Science and Technology and Professor of Applied Physics and is the Director of the Lee Center for Advanced Networking.

Read more at: http://www.vahala.caltech.edu

References

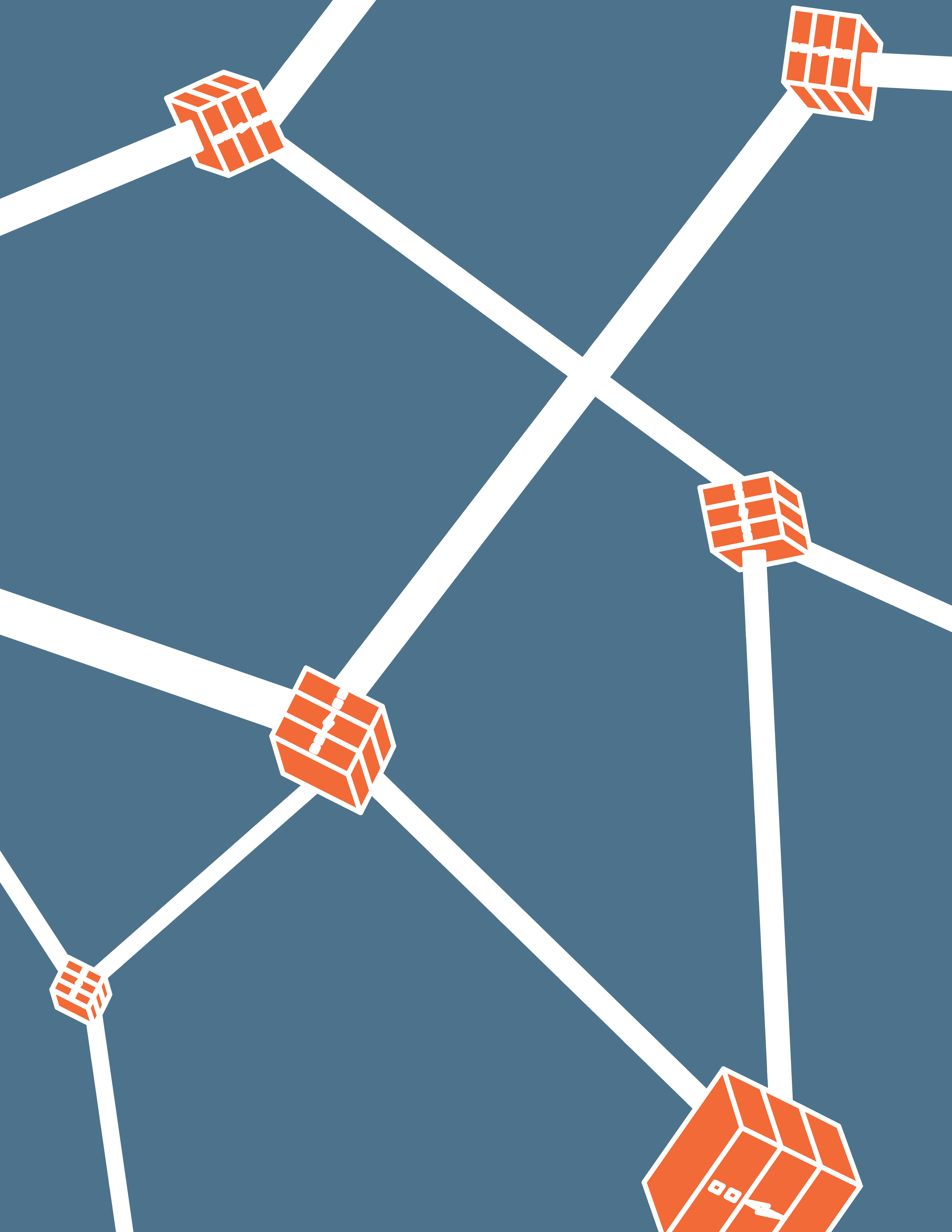
[1] K. J. Vahala, "Optical Microcavities," *Nature*, **424**, 6950 (2003).

[2] D. K. Armani, T. J. Kippenberg, S. M. Spillane and K. J. Vahala, "Ultra-high-Q toroid microcavity on a chip," *Nature,* **421,** 925-929 (2003).

[3] T. Carmon and K. J. Vahala, "Visible continuous emission from a silica microphotonic device by third-harmonic generation," *Nature Physics*, **3**, 430-435 (2007).

[4] T. J. Kippenberg, S. M. Spillane, and K. J. Vahala, "Kerr-nonlinearity optical parametric oscillation in an ultra-high-Q toroid microcavity," *Physical Review Letters*, **93**, 083904 (2004).

[5] H. Rokhsari, T. J. Kippenberg, T. Carmon, and K. J. Vahala, "Radiation-pressure-driven micro-mechanical oscillator," *Optics Express*, **13**, 14 (2005).



Better Network Design Through Measurement and Modeling

by Adam Wierman

Designing and managing networked systems relies on a balance of experimental and analytic tools. Despite the importance of both, recently, experimentation has begun to dominate. Tools like scheduling and queueing theory, which proved invaluable during the development of the internet, are now considered passé. There are many reasons for this shift. The growing complexity of systems makes developing tractable, accurate models increasingly difficult. Further, as we develop a better understanding of system workloads and user behavior, traditional assumptions in models are increasingly being invalidated. Additionally, the traditional metrics studied in scheduling and queueing theory are secondary in importance to measures like power usage, quality of service, and fairness. These changes have resulted in new design paradigms, that no longer fit existing models, e.g. the increasing adoption of multi-channel wireless and multi-core chips to combat power constraints.

> "restore balance between experimental and analytic design in networked systems."

The research in my group seeks to energize analytic performance evaluation, and thus restore balance between experimental and analytic design in networked systems. In order to accomplish this, tools from scheduling and queueing theory need to be "modernized" for today's computer systems. This research requires working closely with both practitioners, in order to understand design decisions being considered, and theoreticians, in order to develop new mathematical tools.

The Impact of the Lee Center

I arrived at Caltech only 2 years ago. However, in this short time the Lee Center has been instrumental in allowing my research to hit the ground running at Caltech. In fact, the Lee Center even played a role in drawing me to Caltech as a result of the financial freedom it provided and the circle of diverse researchers it brings together. The support from the Lee Center allowed me to quickly build my group and also start new research directions without concern as to when external funding would follow. As a result, in only two years, we have had important research successes in projects that began only after I arrived at Caltech.

Research Successes

Lee center funding was essential for each of these projects. Each began during my initial days at Caltech and benefited greatly from the interactions with other faculty and students in the workshops and seminars organized through the Lee Center.

The science of Green IT: The rapidly increasing power consumption of networked systems is a pressing concern for many reasons, including high operating costs, limited battery lifetimes, and a growing carbon footprint. These concerns arise in devices as small as chips, where power limitations have pushed the industry to adopt multi-core designs, all the way to systems as large as data centers, where energy costs are often 40% of the operating budget. As a result, there is a push both in academia and in industry to develop more energy efficient designs.

For many years, the maxim of system design was "faster is better", but now that energy efficiency is important, the maxim has changed to "speed costs power"—there is a tradeoff that must be made between "faster" (smaller delay, larger throughput) and "greener" (less energy). Across all levels of computer systems this tradeoff is typically made via speed scaling, i.e., controlling the speed of the system so as to balance energy and delay (e.g. running slower when fewer jobs are waiting). Speed scaling designs are not new and have been applied for many years in wireless devices and chips; however, the fundamental tradeoffs and limitations of speed scaling designs are not understood.

Our work seeks to explore these tradeoffs analytically and has exposed some important new insights:

- *What is the optimal speed scaler?* We have proven that it is impossible for an online speed scaling design to be optimal across all workloads. Over the past decade analytic research has sought to provide near optimal speed scaling algorithms. Our work proposes an algorithm that improves the best known performance guarantee (our algorithm is 2-competitive) and, further, proves that no online algorithm can be better than 2-competitive. Thus, our results show that scheduling for energy and delay is fundamentally harder than scheduling for delay alone (since it is possible to be optimal for mean delay).

- *How sophisticated must a speed scaler be?* We have proven that a simple speed scaling scheme that sleeps when the system is idle and otherwise runs at a constant speed is nearly as good as the optimal speed scaling scheme, which can dynamically adjust speeds at any point of time. However, the optimal scheme provides a different benefit: robustness, e.g., to time-varying workloads.

- *Does speed scaling have any unexpected consequences?* We have proven that speed scaling increases the unfairness of scheduling policies: large jobs are more likely to be in the system when the server speed is slow. However, this unfairness can be countered by paying a small price of increased energy usage, e.g., via increased speeds at low occupancies.

- *How does scheduling interact with speed scaling?* Our results show that in many cases decisions about processing speed and scheduling order can be decoupled with little loss in performance. Thus, optimal speed scaling algorithms can be determined largely independently of the scheduling policy even though it seems that these decisions are highly intertwined.

“For many years, the maxim of system design was ‘faster is better,’ but now that energy efficiency is important, the maxim has changed to ‘speed costs power’”

Non-cooperative cooperative control: Decentralized distributed resource allocation is an increasingly common paradigm across computer networks. Indeed, it is the dominant paradigm in wireless networks, where centralized control is typically impossible, e.g., for access point assignment, power control, and frequency selection problems. The design of distributed protocols for these problems is difficult and typically protocols come with few analytic guarantees. Our work proposes a (non-cooperative) game-theoretic approach to resource allocation that provides general application-independent techniques for developing efficient distributed designs. In particular, our approach designs the decentralized agents as self-interested players in a game, and then engineers the rules of the game in a way that ensures the equilibria of the game (the stable points) are efficient. When taking such an approach, the key engineering decisions are (i) how to design the rules of the game and (ii) how to design the agents that play the game. Our work gives application independent design rules for each of these decisions. Further, we have developed applications of these techniques, including the sensor coverage problem, network coding, power control in wireless networks, and the access point assignment problem.

Tails of scheduling: Traditional scheduling analysis focuses on performance metrics such as mean delay and mean queue length, while providing little insight into the design of scheduling policies that optimize the distribution of delay and queue length. However, modern networked systems seek quality of service (QoS) measures that depend on distributional characteristics not just expected values. To bridge this gap, we are developing analytic tools to study the distributional behavior of scheduling policies in general settings. We have succeeded in analyzing the distributional behavior of delay under a wide array of common policies in very general settings. From this work has emerged some interesting insights. For example, policies that perform well under light-tailed job sizes perform poorly under heavy-tailed job sizes and vice versa. For ten years, it has been widely conjectured that it is impossible to be optimal for the delay distribution in both light-tailed and heavy-tailed settings. We resolved this conjecture by proving that, indeed, no policy that does not know the job size distribution can be optimal in both settings—and further that if a policy is optimal in one setting it must be worst-case in the other setting. Surprisingly, however, we have also shown that if the policy has very little information about the job size distribution—just its mean—then this is already enough to allow the policy to be near-optimal in both regimes.

Each of these projects are in the initial stages and will continue to produce new insights and designs in the coming years. For each project, the Lee Center provided the initial support that allowed the project to develop the first few results, at which point it became possible to attain external funding to support the project. Now, each projects is self-supporting via government and industrial grants. Thus, the initial support from the Lee Center will continue to resonate for years to come. LEE

Adam Wierman is Assistant Professor of Computer Science.

Read more at: http://www.cs.caltech.edu/~adamw

References

[1] A. Wierman and B. Zwart, "Is Tail-Optimal Scheduling Possible?" Under submission.

[2] J. Marden and A. Wierman, "Overcoming the Limitations of Game-Theoretic Distributed Control," Proceedings of the Conference on Decision and Control (CDC), 2009.

[3] H-L. Chen, J. Marden, and A. Wierman, "On The Impact Of Heterogeneity And Back-End Scheduling In Load Balancing Designs," Proceedings of INFOCOM, 2009.

[4] L. Andrew, A. Wierman, and A. Tang, "Power-Aware Speed Scaling in Processor Sharing Systems," Proceedings of INFOCOM, 2009.

[5] J. Marden and A. Wierman, "Distributed Welfare Games with Applications to Sensor Coverage," Proceedings of the Conference on Decision and Control (CDC), 2009.

Optoelectronic Phase-lock Techniques for the Phase and Frequency Control of a Semiconductor Laser

by Amnon Yariv

Almost all of the information-related applications of lasers to date have been based on a manipulation of the amplitude of the laser radiation. This is in sharp contrast to the field of radio-frequency (RF) electronics where the phase of the radio wave plays a key role. Specifically, phase-lock loop (PLL) systems are the main enablers of many applications—wireless communication, CDMA, FM demodulation and clock recovery, to name a few.

The semiconductor laser (SCL) is the transmission source in optical communication networks, and has a number of unique properties, such as its very large current-frequency sensitivity, fast modulation response, small volume and compatibility with electronic circuits. Research in our group has focused on using these unique properties to import to optics and optical communication many of the important applications of RF electronics. This phase control, when combined with the wide bandwidth of optical waves, can enable a new generation of photonic and RF systems. With support from the Lee center, we are studying many novel applications of optoelectronic phase and frequency control. These applications include: distributed low-cost sensor networks; high-power electronically steerable optical beams; arbitrary waveform synthesis; and wideband, precisely-controlled swept-frequency laser sources for 3-D imaging, chemical sensing and spectroscopy.

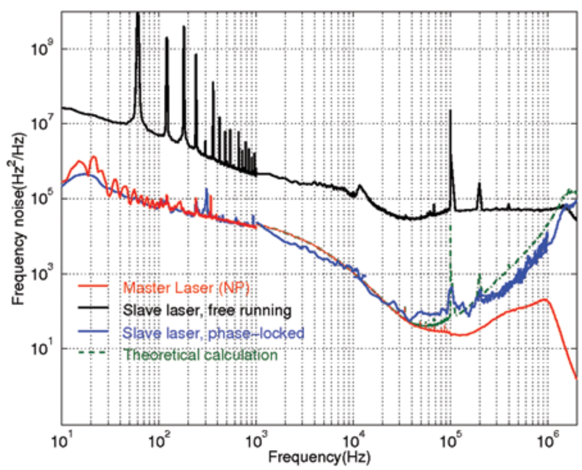


Figure 1: Measured spectrum of the frequency noise of the laser. The frequency noise of the master laser is cloned onto the phase-locked slave laser within the loop bandwidth.

Optical Phase-lock Loops

Optical phase-lock loops (OPLLs) are negative feedback control systems where a SCL acts as a "slave" oscillator whose phase and frequency are locked to the phase and frequency of a high-quality "master" laser. The main differences between an OPLL and electronic PLLs are the wide linewidth and non-uniform frequency tuning characteristics of typical SCLs as compared to electronic oscillators. We have developed novel phase-locking architectures to overcome these effects and achieve stable phase-locking. When the loop is in lock, the phase coherence of the high quality master laser is transferred to the slave laser. This "coherence cloning" is shown in Figure 1, where the frequency noise of the slave laser is reduced by about two orders of magnitude within the loop bandwidth[1]. The improved coherence enables the replacement of expensive narrow linewidth lasers by inexpensive, compact and efficient coherence-cloned SCLs in long range sensing networks.

We have also pioneered the application of phase-locked lasers to phase-controlled coherent arrays for power combination and beam steering. Individual SCLs when locked to a common master laser are coherent relative to each other, and their outputs can therefore be coherently combined. The resulting combined output light from such an array of locked lasers gives a higher peak intensity as compared incoherent combination. Further, we have shown that the use of an RF offset in the OPLL permits electronic control over the optical phase on a one-to-one basis, enabling electronic beam steering and adaptive optics at high speeds.

Broadband Swept Frequency Sources for High Resolution Imaging and Spectroscopy

Swept-frequency—or "chirped"—optical sources with large frequency excursions would be useful in high-resolution 3-D imaging applications. For example, applications such as laser radar, biometrics and optical imaging would benefit greatly from such a source since the depth resolution is inversely proportional to the frequency excursion of the source. A state-of-the-art, chirped-laser source consists of an external cavity laser

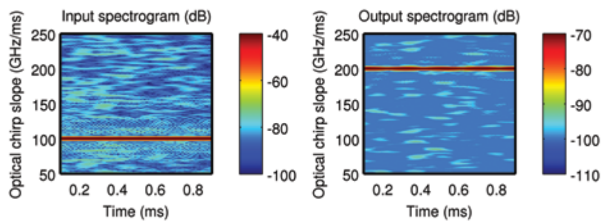


Figure 2: Experimental demonstration of a precisely linear optical frequency chirp using an SCL in an optoelectronic control loop (left) and the doubling of the frequency chirp slope by the process of four wave mixing (right).

whose cavity length, and hence the lasing frequency, is tuned mechanically. The laser is therefore bulky and unreliable; and the tuning speed is limited by the mechanical nature of the tuning mechanism. Further, it is difficult to precisely control the shape of the laser frequency vs. time curve.

We have developed an optoelectronic feedback technique for precise control over the shape of the laser chirp[2], where the chirp is determined by the frequency of an external RF oscillator, and not by the individual laser used in the source. The result is a compact, solid-state laser source with no moving parts that outputs a wideband (>100 GHz), linear frequency chirp vs. time, (see left panel of Figure 2). We have further shown that the slope and bandwidth of the frequency chirp can be doubled by the non-linear optical process of four-wave mixing (FWM)[3], (see right panel of Figure 2). A cascaded implementation of FWM stages can be used to geometrically increase the chirp bandwidth. Moreover, we have demonstrated that a number of chirped SCL sources that chirp over distinct regions of the optical spectrum can be combined to effectively approximate a single wide-bandwidth chirp[4], an approach similar to techniques employed in synthetic aperture radar.

Ongoing and Future Work

With the advent of hybrid III-V/silicon lasers and photonic integrated circuits on silicon and III-V platforms, our future research efforts will focus on the integration of functions (see Figure 3).

Beyond the applications mentioned earlier, we are also exploring use of electronically controlled chirped sources in spectroscopy, and chemical and biological sensing, where the narrow linewidth and precisely controlled frequency of the optical source enables stable, high-resolution frequency measurements. Further, by mixing a chirped laser with another optical wave on a high-speed mixer, these broadband sources can serve as sources in terahertz spectroscopy and imaging.

Amnon Yariv is the Martin and Eileen Summerfield Professor of Applied Physics and Professor of Electrical Engineering.

Read more at:

http://www.its.caltech.edu/~aphyariv/base.html

References

[1] N. Satyan, W. Liang, and A. Yariv, "Coherence Cloning using Semiconductor Laser Optical Phase-lock Loops," *IEEE J. Quantum Elect*. **45**, 755–761 (2009).

[2] N. Satyan, A. Vasilyev, G. Rakuljic, V. Leyva, and A. Yariv, "Precise Control of Broadband Frequency Chirps using Optoelectronic Feedback," *Opt. Express* **17**, 15991–15999 (2009).

[3] N. Satyan, G. Rakuljic, and A. Yariv, "Chirp Multiplication by Four Wave Mixing for Wideband Swept-frequency Sources for High Resolution Imaging," (manuscript in preparation).

[4] A. Vasilyev, N. Satyan, S. Xu, G. Rakuljic, and A. Yariv, "Multiple-source FMCW reflectometry: theory and experiment," *Appl. Opt.*, in press.

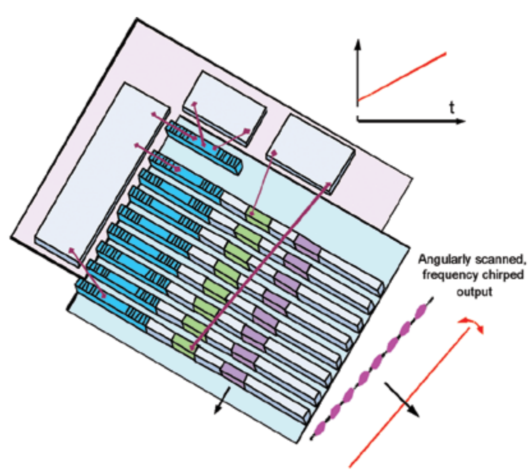


Figure 3: Schematic diagram of an integrated, frequency-chirped coherent optical source.

Wavestream Corporation

New Power for Satellite Systems

by David Rutledge

The Wavestream Corporation manufactures high-power solid-state transmitters. The company is based on research developments in David Rutledge's group at Caltech with funding from the Lee Center for Advanced Networking and the Army Research Office.

The 30-GHz band is an attractive frequency range for high-speed satellite uplinks, except, historically, the performance has been limited by the low output power of traditional integrated circuits. The key challenge has been high losses in the on-chip transmission-line circuits that combine the outputs from the transistors. At Caltech, we conceived the idea of combining the power from transistors in the air above the chip. This approach, called active quasi-optics, eliminates the transmission-line losses, and allows much higher power levels than was possible before. Wavestream's best-selling product has an output power of 50 Watts between 30 GHz and 31 GHz. For comparison, previous circuits were limited to 5 Watts. The increase in power provides a direct 10:1 improvement in data transmission rates. In addition, the amplifier is light enough to be mounted directly at the focus of the reflector antenna, eliminating further losses in the connecting waveguide.

Wavestream is based in San Dimas. It has 130 employees, and last year the company had 50 million dollars in sales. Former Caltech graduate students and post-doctoral fellows are prominent in the company. Michael DeLisio, a founder, is the Chief Technical Officer. Chad Deckman, also a founder, is Vice President for Research. IST postdoctoral fellow Younkyu Chung recently joined the company. Laurence Cheung is the head of the Wavestream Singapore office.

David B. Rutledge is the Kiyo and Eiko Tomiyasu Professor of Electrical Engineering.

Read more at: http://www.wavestream.com

"we conceived the idea of combining the power from transistors in the air above the chip...called active quasi-optics"

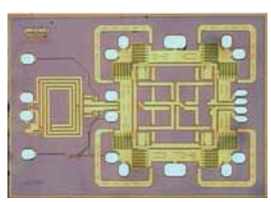
CMOS Power

Axiom Microdevices, Inc.

by Ali Hajimiri

CMOS integrated circuits took over the area of wireless communications in the late 90's. This lead to creation of giants such as Broadcom which executed the fab-less semiconductor model and managed to transform wireless transceiver design from somewhat of a "black magic" art to a systematic process. Despite the success of such players, RF power generation and amplification remained the realm of compound semiconductor devices in conjunction with a PCB module to integrate the necessary passive elements. Many in the field believed it impossible to create an integrated CMOS power amplifier. Such assertions were also supported with calculations based on classical amplified topologies and CMOS transistor parameters.

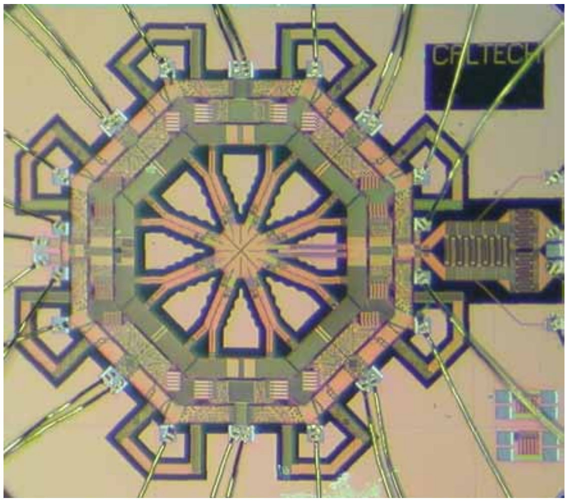
The Caltech Lee center gave us a unique opportunity to challenge this universally-held dogma by going after the most challenging problem out there; namely, the fully integrated CMOS PA (power amplifier). It created an environment where two star students (Dr. Ichiro Aoki and Dr. Scott Kee) could work in two groups with complementary sets of skills in silicon integrated circuit design (Hajimiri) and microwave systems (Rutledge). The result was magic: we came up with a truly innovative and fundamentally different technique called the distributed active transformer (DAT). The first prototype of a fully-integrated, CMOS PA generated more than 2 watts of RF power at 2.4GHz without any external components. It became clear that a large market, previously the realm of module-based compound semiconductors, could be addressed using this innovation. We chose to go after the cellular phone market with more than 1 Billion new phones sold each year.



CMOS power amplifier

This decision led to the formation of Axiom Microdevices Inc. in 2002 which successfully took the original Caltech technology and developed it much further to overcome many important performance and market challenges before product launch. Axiom successfully started the production run of AX502, the world's first (and to this date only) fully-integrated CMOS power amplifier for GSM/GPRS/EDGE, in February of

"The result was magic: we came up with a truly innovative and fundamentally different technique called the distributed active transformer"



A distributed active transformer (DAT) design

"millions of people around the world have access to smaller, more efficient phones that cost less and last longer"

2007; and quickly ramped production so that by June that year it was at the 1 million parts shipped mark. The ramp-up continued, and the company shipped more than 50 Million parts in May 2009, when Axiom was acquired by Skyworks Inc. Skyworks has continued to produce the parts and expand the product portfolio, shipping, on average, 3 Million parts per month.

Axiom has been a true success story of taking a high-risk idea from academia to commercial product. Lee Center funding played an important role in seeding the early research in our laboratory, precisely at a time when such funding was most difficult to obtain from conventional sources. Because of this unique opportunity to go after the most challenging problems out there, millions of people around the world have access to smaller, more efficient phones that cost less and last longer. Thank you Ellen and David! L E E

Ali Hajimiri is Thomas G. Myers Professor of Electrical Engineering.

Read more at: http://www.axiom-micro.com

The Aerospace Corporation in the Lee Center

by Paul Anderson and Fletcher Wicker

Participation in Lee Center activities has provided a clear benefit to The Aerospace Corporation. The topic of communication networks in space is of significant interest to our government customers. Knowledge of current research topics, interaction with Caltech faculty and students, and Caltech faculty exposure to significant space network design issues have all helped to advance the use of communication networks in national security space programs.

AEROSPACE
Assuring Space Mission Success

The Aerospace Corporation operates a federally funded research and development center (FFRDC) that is sponsored by the United States Air Force. FFRDCs are unique, independent nonprofit entities that are funded by the U.S. government to meet specific long-term technical needs that cannot be met by any other single organization. FFRDCs typically assist government agencies with scientific research and analysis, systems development, and systems acquisition. They bring together the expertise and outlook of government, industry, and academia to solve complex technical problems.

For national security relating to space, The Aerospace Corporation supports long-term planning and the immediate needs of our nation's military and reconnaissance space programs. The Aerospace Corporation provides scientific and engineering support for launch, space, and related ground systems. It also provides the specialized facilities and continuity of effort required for programs that often take decades to complete. This end-to-end involvement reduces development risks and costs, and allows for a high probability of mission success. The primary customers are the Space and Missile Systems Center of Air Force Space Command and the National Reconnaissance Office, although work is performed for civil agencies as well as international organizations and governments in the national interest.

“unique partnerships”

Within the company, The Aerospace Institute coordinates interaction with universities and colleges such as Caltech, and the Corporate University Affiliates Program was initiated in 1997 to create formal relationships between The Aerospace Corporation and selected universities. Participating universities and academic departments engage in formal agreements with The Aerospace Corporation to accomplish specific tasks and share technical information. The agreements create unique partnerships between faculty, students, and The Aerospace Corporation technical staff as they work together to exchange technology in critical skill areas. In addition, there is an Informal University Relations Program that provides support for employee participation in ad hoc activities at local colleges and universities. These informal activities may include faculty collaborations, technical seminars/colloquia, student mentoring, research reviews, etc. It was through these programs that The Aerospace Corporation became aware of the Caltech Lee Center. Participation in Lee Center activities began in the fall of 2006 and continued until the termination of Lee Center activities in the spring of 2009.

Interest in networking technologies has dramatically increased in recent years across many national security space programs. Of particular interest is the application of these technologies to space-based networks. Participation by The Aerospace Corporation in the Lee Center seminars and workshops provided (1) an understanding of current network-related research in academia to better support the company's customers and (2) an interaction with new Caltech graduates for possible employment at the company, either through direct contact with these students or through recommendations by Caltech faculty. It has also exposed Caltech faculty to current network design and performance issues of interest to The Aerospace Corporation.

Specific benefits to The Aerospace Corporation due to past participation in Lee Center activities have included interactions with the following professors at Caltech: Professor Hajimiri on novel methods for modulating signals through antenna switching, Professor Wierman on optimization of queuing systems, Professor Effros on network capacity theory, Professor Yariv

"space-based networks"

on optoelectronic devices, and Professor McEliece on information theory and coding. In addition, the sponsorship of the 2008 summer undergraduate research fellowship (SURF) for student Arthur Chang was inspired by Lee Center interactions, as was the company's 2009 hiring of engineering intern Daniel Thai, who was a company-sponsored Caltech SURF fellow and whose interests are in network theory. Moreover, on May 6, 2009 Professor Steven Low of Caltech visited The Aerospace Corporation for discussions of possible research collaboration on the use of fast TCP protocols over satellite links.

Dr. Fletcher Wicker, Dr. Mark Coodey, Mas Sayano, and Paul Anderson of The Aerospace Corporation attended many of the lunchtime seminars sponsored by the Lee Center. Dr. Wicker and Paul Anderson also presented material on two of these occasions. In the fall term of 2006, Dr. Fletcher Wicker discussed quality of service issues for space-based routers serving downlinks with time-varying data rates, and in the winter term of 2009, Paul Anderson discussed satellite sizing and capacity trends and the use of networks in the commercial satellite industry. LEE

Dr. Paul Anderson (left) and Dr. Fletcher Wicker (right); The Aerospace Corporation.

Read more at: http://www.aero.org

Lee Center & the IST Initiative

Vision for Caltech

by Jehoshua (Shuki) Bruck and Richard Murray

The Information Science and Technology (IST) initiative began in 2001–02 as part of the strategic planning activities within the Division of Engineering and Applied Science. A faculty committee, led by Shuki Bruck, was formed to establish a vision for Caltech's activities in this important area. The Lee Center provided a strong community building mechanism that connected between faculty in Electrical Engineering, Computer Science, Physics and Economics and served as a fertile ground for the hatching, evolution and definition of the IST initiative. From the earliest discussions, the Lee Center served as a model for how to organize the research activities within IST. In particular, the Lee Center's emphasis on providing flexible resources to faculty that allowed them pursue new research directions prior to receiving external grants was seen as an exceptional feature that should be emulated within IST.

Based on the recommendations of the IST faculty committee, the IST initiative was formally launched in 2003–04. It combined two existing centers, the Lee Center and the NSF Center for Neuromorphic Systems Engineering (CNSE), with four new centers: the Center for the Physics of Computation (CPI), the Center for Biological Circuit Design (CBCD), the Social and Information Science Laboratory (SISL) and the Center for the Mathematics of Information (CMI). These centers allowed research in a number of new areas, including networking principles in social organizations; storage; processing and retrieval of information in biological structures; technologies based on non-silicon computational material including approaches to quantum computing and the study of mathematical principles that underlie all of these activities.

PHYSICS

Professor Shuki Bruck served as the initial Director of IST, and a steering committee of 8 members put into place the organizational structure to move this enterprise forward. Over the last 7 years, several faculty searches were carried out to identify new faculty hires related to IST, including a broad search and several more focused searches in biological circuits, computer science, economics and electronics.

BIOLOGICAL CIRCUITS

Each of the new centers in IST has catalyzed a set of activities that have moved Caltech forward in new research directions:

CENTER FOR THE PHYSICS OF INFORMATION

Information is something that can be encoded in the state of a physical system, and a computation is a task that can be performed with a physically realizable device. Thus the quest for better ways to acquire, store, transmit, and process information leads us to seek more powerful methods for understanding and controlling the physical world. Limitations inherent in the physical form of information (such as the size of atoms and the quantum effects that unavoidably arise in very small systems) pose great challenges that must be overcome if information technology is to continue to advance at the rate to which we have grown accustomed. The CPI is dedicated to the proposition that physical science and information science are interdependent and inseparable. Its research aims to foster physical insights that can pave the way for revolutionary new information technologies and to stimulate new ideas about information that can illuminate fundamental issues in physics and chemistry.

CENTER FOR BIOLOGICAL CIRCUIT DESIGN

The Center for Biological Circuit Design is developing new ways to design, build and analyze biological circuits. These circuits control information flow in biological systems, and as such are a core area of IST. The study of circuits cuts across vast areas of biology, from biochemistry, biophysics and genetics, to cell and developmental biology, to neurobiology and ecology. Understanding how to design and build circuits is crucial for the next generation of bioengineering. The study of biological circuits also opens up new areas for theory in computation. Research in the CBCD combines the experimental biologist's desire to abstract the key principles from the richness and diversity of biological circuits, the physicist's sense of measurement and of simple underlying mechanisms, and the engineer's aesthetic of "to build is to understand." The CBCD is an interdisciplinary group of biologists and engineers from a broad range of engineering and biology disciplines.

SOCIAL AND INFORMATION SYSTEMS LABORATORY

The Social and Information Sciences Laboratory (SISL) studies how markets and other social systems aggregate large amounts of information that is widely distributed. Researchers in SISL are also working to design new and improved markets, network protocols, sensor systems, and political processes. Some of the specific topics being investigated by SISL researchers include the design of combinatorial auctions used for privatization; the design of large-scale interactive distributed systems such as electricity markets; the study of price formation and the possibilities for use of market-based systems for information aggregation in a variety of settings; the optimal structuring of elections, committees, and juries; the formation and evolution of networked systems with independent actors; and many of the computational aspects of system design (e.g., what do users need to compute and what does the system need to compute?).

SOCIAL AND INFORMATION SYSTEMS

MATHEMATICS

One of the novel aspects of SISL research concerns understanding how humans interact with technology and what that implies about the design of the technology. Since such systems involve both human behavior and technology, SISL brings together researchers from the social sciences, engineering, and applied and computational mathematics.

CENTER FOR THE MATHEMATICS OF INFORMATION

The ability to measure and model physical quantities—energy, temperature, and the like—has inspired the development of modern mathematics. Scientific and engineering questions have provided a source of vitality to, among other fields, algebra, geometry and analysis. Pure mathematical research in those fields coexists today with mathematic's function as a language and tool for the physical sciences. A concern with logical and statistical quantities—information—is central to present-day science and engineering. Indeed, disciplines such as algorithms, complexity, communications and control motivate questions in adjoining fields such as probability, combinatorics, algebra and harmonic analysis. A common framework supporting the study of information and computation across disciplines is as yet a distant goal. The Caltech CMI is a home in which unfettered development of the mathematical foundations of information and computation can be influenced by, and influence in turn, progress in engineering and science. These centers have substantial overlap with each other and with the Lee Center, allowing exchange of ideas across different scientific and engineering applications. L E E

Jehoshua (Shuki) Bruck (left)is the Gordon and Betty Moore Professor of Computation and Neural Systems and Electrical Engineering, and Richard Murray (right) is the Thomas E. and Doris Everhart Professor of Control & Dynamical Systems and Bioengineering.

Read more at: http://www.ist.caltech.edu

2007 Urban Grand Challenge

by Richard Murray

The Lee Center was one of the funders of Caltech's entry in the 2007 Urban Grand Challenge, an international autonomous vehicle competition sponsored by DARPA. Team Caltech was formed in February of 2003 with the goal of designing a vehicle that could compete in the 2004 DARPA Grand Challenge. Our 2004 vehicle, Bob, completed the qualification course and traveled approximately 1.3 miles of the 142-mile 2004 course. In 2004-05, Team Caltech developed a new vehicle, Alice, to participate in the 2005 DARPA Grand Challenge. Alice utilized a highly networked control system architecture to provide high performance, autonomous driving in unknown environments. The system completed several runs in the National Qualifying Event, but encountered a combination of sensing and control issues in the Grand Challenge Event that led to a critical failure after traversing approximately 8 miles.

As part of the 2007 Urban Challenge, Team Caltech developed new technology for Alice in three key areas:

1. mission and contingency management for autonomous systems
2. distributed sensor fusion, mapping and situational awareness
3. optimization-based guidance, navigation and control.

Support from the Lee Center was used to developed some of this technology, primarily through support for undergraduate research projects (SURFs) and the purchase of system hardware.

For the 2007 Urban Challenge, we built on the architecture that was deployed by Caltech in the 2005 race, but provided significant extensions and additions that allowed operation in the more complicated (and uncertain) urban driving environment. Our primary approach in the desert competition was to construct an elevation map of the terrain surrounding the vehicle and then convert this map into a cost function that could be used to plan a high-speed path through the environment. A supervisory controller provided contingency management by identifying selected situations (such as loss of GPS or lack of forward progress) and implementing tactics to overcome these situations.

For the urban environment, several new challenges had to be addressed. These included the determination of road location based on lane and road features; avoidance of static and moving obstacles; and successful navigation of intersections. We chose a deliberative planning architecture, in which a representation of the environment was built-up through sensor data and motion planning was done using this representation. A significant issue was the need to reason about traffic situations in which the vehicle interacts with other vehicles

Alice

Top row (L to R): unknown (DARPA), Sam Pfister, Noel duToit, Tony Tether (former DARPA director), Cisco Zabala, Tony Fender, Mohamed Aly, Humberto Pereira, Andrew Howard (JPL), Richard Murray, Laura Lindzey, Sandie Fender. Bottom row (L to R): Nok Wongpiromsarn, Jeremy Ma, Josh Oreman, Christian Looman.

or has inconsistent data about the local environment or traffic state.

The following technical accomplishments were achieved as part of this project:

1. A highly distributed, information-rich sensory system was developed that allowed real-time processing of large amounts of raw data to obtain information required for driving in urban environments. The distributed nature of our system allowed easy integration of new sensors, but required sensor fusion in both time and space across a distributed set of processes.

2. A hieararchical planner was developed for driving in urban environments that allowed complex interactions with other vehicles, including following, passing and queuing operations. A rail-based planner was used to allow rapid evaluation of maneuvers and choice of paths that optimized competing objectives while also insuring safe operation in the presence of other vehicles and static obstacles.

3. A canonical software structure was developed for use in the planning stack to insure that contingencies could be handled and that the vehicle would continue to make forward progress towards its goals. The combination of a directive/response mechanism for intermodule communication and fault-handling algorithms provide a rich set of behaviors in complex driving situations.

The features of our system were demonstrated in over 300 miles of testing performed in the months before the race, including the first known interaction between two autonomous vehicles (with MIT, in joint testing at the El Toro Marine Corps Air Station)

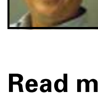
A number of shortfalls in our approach led to our vehicle being disqualified for the final race:

1. Inconsistencies in the methods by which obstacles were handled led to incorrect behavior in situations with tight obstacles.

2. Inadequate testing of low-level feature extraction of stop-lines created problems with the corresponding fusion into the existing map.

3. Adjustments needed for handling intersections and obstacles proved difficult to modify and test in the qualification event.

Despite these limitations in our design, Alice was able to perform well on 2 out of the 3 test areas at the National Qualifying Event, demonstrating the ability to handle a variety of complex traffic situations. Most important, a new generation of researchers was able to participate in the design, implementation and testing of one of the handful of completely autonomous vehicles that have been developed.



Richard Murray is the Thomas E. and Doris Everhart Professor of Control & Dynamical Systems and Bioengineering.

Read more at: http://www.cds.caltech.edu/~murray

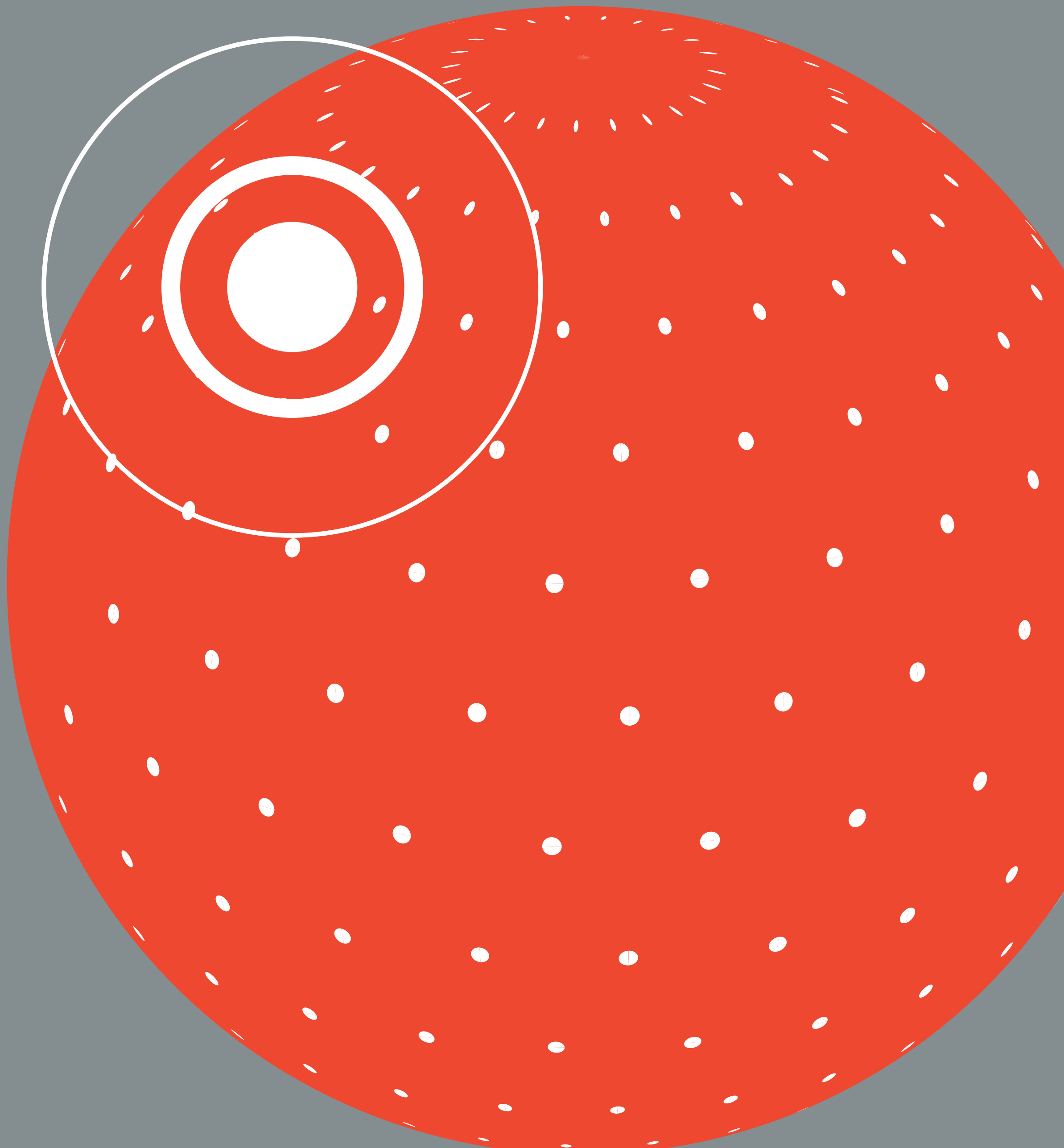
Produced by the Lee Center

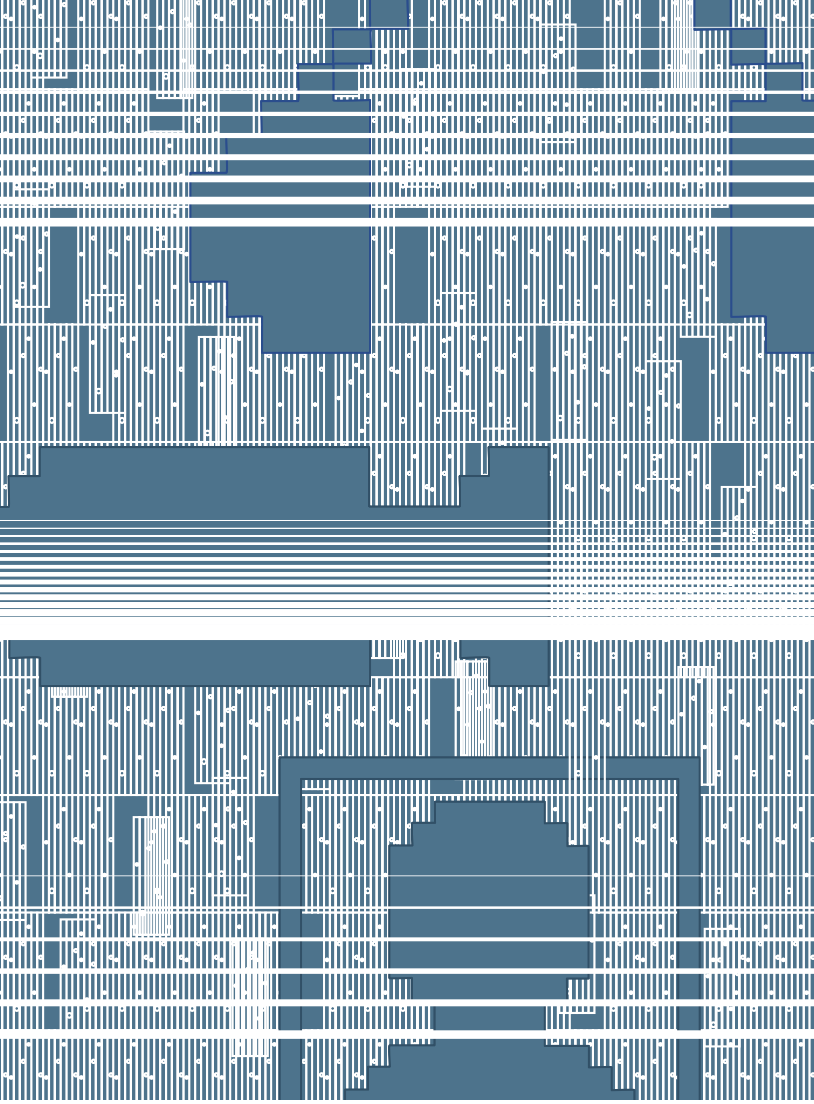
Editors
Kerry Vahala
Trity Pourbahrami

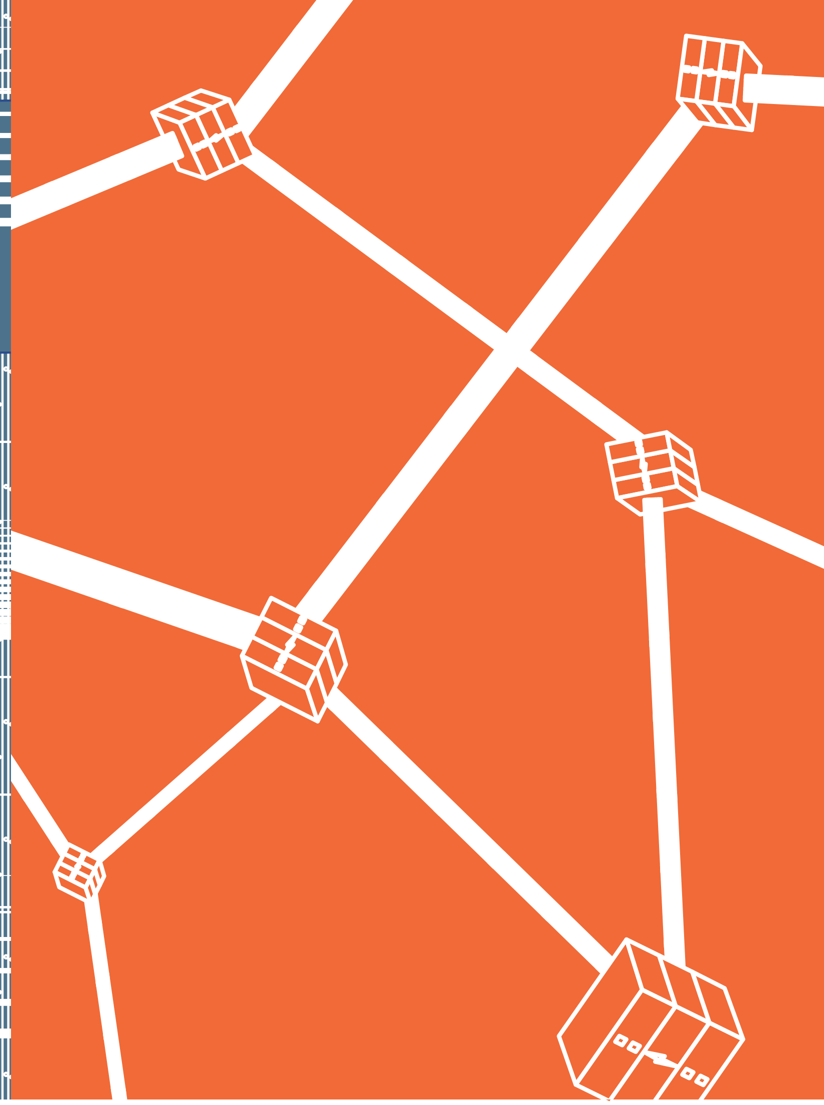
Principal Photography
Peter Mendenhall

Illustrations
John Hersey

Design
Engineering & Applied Science Communications Office







Lee Center
MS 136-93
California Institute of Technology
Pasadena, CA 91125