Participatory Sensing: Applications and Architecture

Deborah Estrin
University of California, Los Angeles

Participatory sensing is the process whereby individuals and communities use ever-more-capable mobile phones and cloud services to collect and analyze systematic data for use in discovery. The convergence of technology and analytical innovation with a citizenry that is increasingly comfortable using mobile phones and online social networking sets the stage for this technology to dramatically impact many aspects of our daily lives.

Applications and Usage Models

One application of participatory sensing is as a tool for health and wellness. For example, individuals can self-monitor to observe and adjust their medication, physical activity, nutrition, and interactions. Potential contexts include chronic-disease management and health behavior change. Communities and health professionals can also use participatory approaches to better understand the development and effective treatment of disease. For some real-world examples, visit www.projecthealthdesign.org and http://your.flowingdata.com.

The same systems can be used as tools for sustainability. For example, individuals and communities can explore their transportation and consumption habits, and corporations can promote more sustainable practices among employees. For examples, visit http://peir.cens.ucla.edu and http://biketastic.com.

In addition, participatory sensing offers a powerful "make a case" technique to support advocacy and civic engagement. It can provide a framework in which citizens can bring to light a civic bottleneck, hazard, personal-safety concern, cultural asset, or other data relevant to urban and natural-resources planning and services, all using data that are systematic and can be validated. For an example, visit http://whatsinvasive.com.

These different applications imply several different usage models. These models range from public contribution, in which individuals collect data in response to inquiries defined by others, to personal use and reflection, in which individuals log information about themselves and use the results for personal analysis and behavior change. Yet across these varied applications and usage models, a common workflow is emerging, as Figure 1 illustrates.

Essential Components

Ubiquitous data capture and *leveraged data processing* are the enabling technical components of these emerging systems. The need for the individual to control access to the most intimate of these data streams introduces a third essential component: the *personal data vault*.

Ubiquitous Data Capture

While empirical data can be collected in a variety of ways, mobile phones are a special and, perhaps, unprecedented tool for the job. These devices have become mobile computing, sensing, and communication platforms, complete with image, audio, video, motion, proximity, and location data capture and broadband communication, and they are capable of being programmed for manual, automatic, and context-aware data capture.

Because of the sheer ubiquity of mobile phones and associated communication infrastructure, it is possible to include people of all backgrounds nearly everywhere in the world. Because these devices travel with us, they can help us make sustainable observations on an intimately personal level. Collectively, they provide unmatched coverage in space and time.

Leveraged Data Processing and Management

In some cases, the data collected with a mobile device are enough to reveal an interesting pattern on their own. However, when processed through a series of external and cross-user data sources, models, and algorithms, simple data can be used to infer complex phenomena about individuals and groups. Mapping and other interactive capabilities of today's Web enhance the presentation and interpretation of these patterns for participants. Many applications will call for the comparison of current measures to past trends, so robust and long-term storage and management of this data is a central requirement.

The Personal Data Vault

A common feature uniting these applications is the highly individualized, and therefore per-

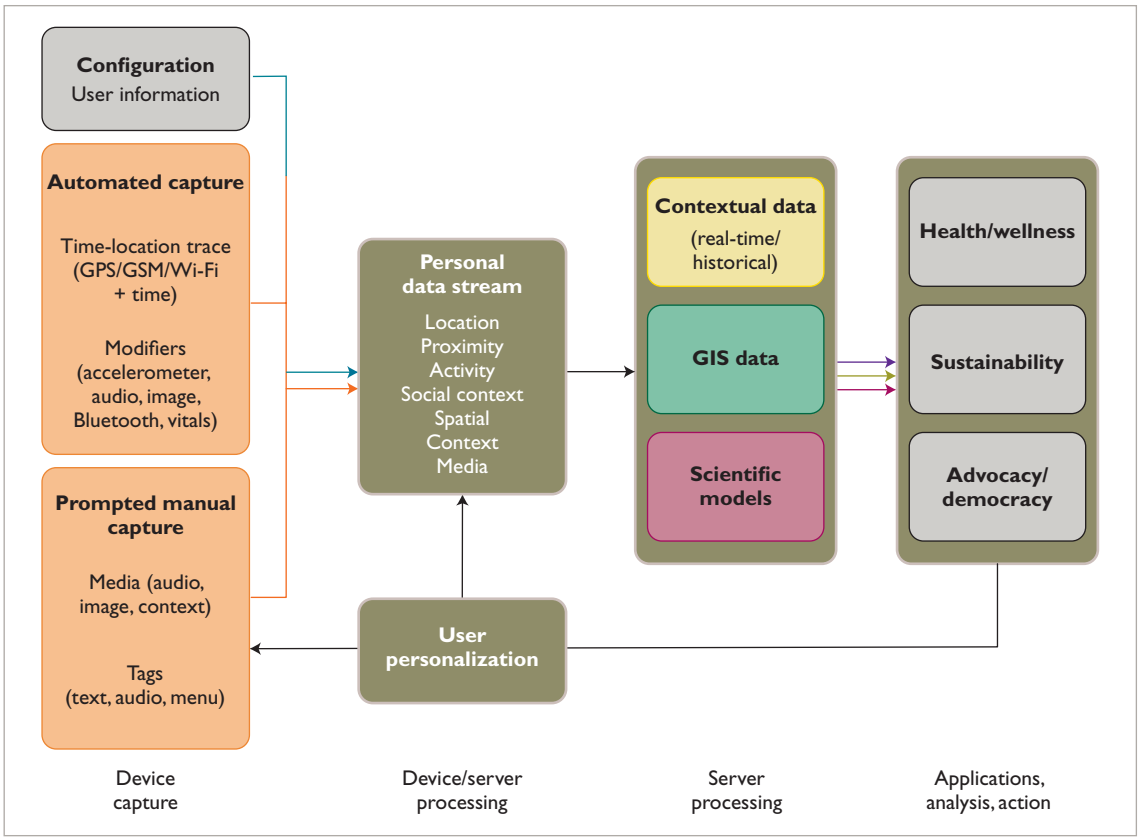


Figure 1. Common architectural components for participatory-sensing applications, including mobile-device data capture, personal data stream storage, and leveraged data processing.

sonal, nature of the data. By building mechanisms for protecting personal data directly into the emerging participatory-sensing architecture, we can create a healthy marketplace of content and services in which the individual has visibility and negotiating power with respect to the use and disposition of his or her personal data streams. By specifying standard mechanisms instead of standard policy, we enable support of diverse policies that are tailored to particular applications and users — this is the *narrow waist* of this participatory-sensing architecture. Without such an architecture, critical applications will be encouraged to create bundled, vertically integrated, non-interoperable, and nontransferable vehicles for personal data streams, thereby making those streams opaque to their creators. By creating such a user-transparent architecture that places individuals and communities at the locus of control over information flow, we will simultaneously support participant rights and create a healthier market for competitive services.

To support this function, we propose the personal data vault. It decouples the capture and archiving of personal data streams from the sharing of that information. Instead of individuals sharing their personal data streams directly with services, we propose the use of secure containers to which only the individual has complete access. The personal data vault would then facilitate the selective sharing of subsets of this information with various services over time. Selective sharing may take the form of exporting filtered information from specific times of day or places in space, or may import service computations to the data vault and export resulting computational outputs. Essential to this scheme are tools to audit information flows and support meaningful usage. Finally, legal consideration is essential to protect and preserve the individual's control over his or her own data streams.

Participatory-sensing systems leveraging mobile phones offer unprecedented observational capacity at the scale of the individual. At the same time, they are remarkably scalable and affordable given the wide proliferation of

cellular phone infrastructure and consumer devices that incorporate location services, digital imagers, accelerometers, Bluetooth access to off-board sensors, and easy programmability. These systems can be leveraged by individuals and communities to address a range of civic concerns, from safety and sustainability to personal and public health. At the same time, they will push even further on our societies' concepts of privacy and private space.

Acknowledgments

This article represents joint work with Jeff Burke, Jeff Goldman, Eric Graham, Mark Hansen, Jerry Kang, Nithya Ramanathan, Sasank Reddy, Mary Jane Rotheram-Borus, Katie Shilton, Mani Srivastava, and the larger urban-sensing group at the Center for Embedded Networked Sensing.

Deborah Estrin is a professor of computer science with a joint appointment in electrical engineering at the University of California, Los Angeles, where she holds the Jon Postel Chair in Computer Networks. She's also the founding director of the university's Center for Embedded Networked Sensing. Her research focuses on scalable, personal, and environmental-sensing applications. Estrin has a PhD in computer science from the Massachusetts Institute of Technology. She's a fellow of the ACM and IEEE and a member of the US National Academy of Engineering. Contact her at destrin@cs.ucla.edu.

The Impact of Sense and Respond Systems

K. Mani Chandy
California Institute of Technology

Sense and respond (S&R) systems based on information technology amplify one of the most fundamental characteristics of life — the ability to detect and respond to events. Living things thrive when they respond effectively to what's going on in their environments. A zebra that doesn't run away from a hungry lion dies and one that runs away unnecessarily wears out. Organizations sense and respond collectively: lions in a pride signal each other when they hunt; societies deal with crises by harnessing capabilities of governments, charities, and individuals. When our ancestors hunted millennia ago, they saw as far as the eye could see and threw spears as far as their muscles let them. Today, S&R systems let us detect events far out in space and respond anywhere on the globe. By 2020, S&R systems will become an integral part of the activities of people and organizations around the world whether they're rich or poor, in farming or medicine, at work or at play.

Mammals sense and respond by using the complementary functions of the sympathetic and parasympathetic nervous systems. The sympathetic nervous system manages "fight or flight" responses while the parasympathetic nervous system handles ongoing functions such as digestion. By 2020, individuals and companies will routinely use S&R systems to amplify their sympathetic and parasympathetic nervous systems: they'll use them to improve the efficiency of day-to-day operational activities and also to respond to rare, but critical, threats and opportunities.

S&R systems have different characteristics than traditional information technology services:

- S&R systems interact with the environment. Computation and communication are relevant only insofar as they support interaction.
- S&R systems direct the activities of components such as sensors, computation engines, data stores, and responders. The programming metaphor for S&R systems is agent choreography rather than the sequential flowchart of a cooking recipe.
- People configure S&R systems to operate over a longer term than conventional service invocations. The invocation of a service — such as a Web search for documents dealing with the keyword "Internet" — handles an immediate, and possibly transient, concern. By contrast, a request to receive alerts about new documents with the keyword "Internet" asks for a longer-term interaction; the request remains in place until the requester deletes it.
- S&R systems are predictive and proactive: they predict what organizations and individuals will need to do, and they carry out activities that users might need in the future. The results of these proactive activities are discarded if the user doesn't need them. A simple example of such a proactive system is one that determines your best commutes to both the office and the airport; if you go to the office, then the work in determin-

ing the optimum commute to the airport is wasted. The decreasing costs of computation and communication compared to the costs of other goods and services will result in more proactive applications.

Feedback control has been widely used since James Watts' centrifugal governor in the 18th century. Militaries have had command and control systems based on information technology since World War II. Market makers in stocks employ complex algorithms that respond to events in milliseconds. Businesses have used intelligence algorithms for more than 25 years. All these technologies are examples of S&R systems. So, what's new about 2020?

S&R technologies will become commonplace in 2020. What was once the exclusive province of sophisticated engineering companies, military contractors, and financial firms will be used by everybody: teenagers, homemakers, senior citizens, CIOs, CFOs, and CEOs. They'll use S&R technologies in 2020 as naturally as they use search engines today.

What forces will make S&R commonplace in 2020?

- Advertising revenues will drive dot-com companies to offer services that allow consumers to create personal S&R systems, including activity-specific dashboards that integrate calendar, mail, Web searches, news alerts, stock feeds, and weather forecasts for aspects ranging from bicycling to investing. Nothing provides more information about you than what you want to sense and how you want to respond, and advertising companies will offer services to gain that data and target "markets of one."
- Decreasing sensor and responder costs and form-factors will drive penetration of S&R systems. Accelerometers that cost hundreds of dollars will cost a tenth as much when they become commodity components of mobile phones and laptops. A rich variety of sensors, such as heart monitors, will be coupled to mobile phones. GPS devices will drive location-based S&R services.
- Programmers will be able to easily structure S&R applications to exploit clusters of machines and multicore computers.
- Advances in several areas of information technology will simplify implementations of S&R systems. These areas include information extraction from natural language text, images, and videos; business intelligence, analytics, machine learning, and optimization; notations and user interfaces for specifying S&R systems; and personal devices such as smart phones and smart clothing.

S&R systems will support all aspects of daily living: water, food, health, energy, security, housing, transportation, and research. Green energy resources such as wind and solar power are dynamic; so, systems that harness these resources must sense their availability and respond appropriately. Indeed, the smart grid can't exist without S&R technologies. Concern about food safety will lead to national farm identification systems that track every farm animal with an RFID tag or microchip. By 2020, many countries will require electronic pedigree systems that record major events — such as shipment and prior sales of pharmaceutical drugs. S&R technologies will play central roles in science projects such as the Large Hadron Collider, and they'll play an even larger role in national defense.

Community-based S&R systems will empower hundreds of thousands of ordinary people equipped with sensors and responders in their mobile phones, cars, and homes to help their communities. People in earthquake zones such as Lima, Jakarta, and Los Angeles will use inexpensive personal accelerometers to send information about ground movement to S&R systems that determine epicenters and provide short (seconds) of warning of intensive shaking. Community-based measurements of wind speed, temperature, and humidity will provide firefighters with microscale data when fighting forest fires in Greece, California, and Australia. Ordinary people will use sensors and the Internet to collaborate on citizen-science projects — for instance, amateur and professional astronomers across the globe working together to record transient astronomical events.

The widespread use of S&R has some dangerous consequences and faces several hurdles:

- An insidious consequence of a badly designed S&R system is that it can dissipate one of the truly scarce resources of this century: attention. Well-designed S&R systems amplify attention, whereas poorly designed

systems dissipate it by interrupting us and giving us opportunities to get sidetracked.
- Concerns about privacy are a barrier. An S&R application will make individuals and organizations more effective; however, the company that hosts the application will know the most important aspect of its users — their goals.
- Security is a major hurdle. Widespread use of sensors and responders gives hackers multiple points of entry into S&R systems. The systems that form the backbone of critical services such as food, water, energy, and finance are likely to have common components; successful attacks or errors in these components will have devastating consequences.
- S&R systems enable efficient use of limited infrastructure, such as electric grids and roads, by distributing demand over time and reducing peak congestion. As a consequence, the infrastructure operates close to capacity much of the time, and an increase in demand can take it over the edge and bring the system down. Resilience requires some spare capacity as well as S&R technology.

Society will feel the impact of S&R technologies in many ways. S&R systems will let people conduct a variety of new services from anywhere. They'll let nurses in Manila monitor senior citizens in Manhattan, and engineers in Bangalore monitor intrusion into buildings and networks in London. S&R technologies will accentuate the digital divide; those who master the technology will function better at school and work than those who don't.

The next 10 years will see rapid development of S&R technologies in applications that touch the daily lives of people across the globe.

K. Mani Chandy is the Simon Ramo Professor at the California Institute of Technology. His research interests include sense and respond systems, event-driven architectures, and distributed systems. He blogs at senseandrespond.blogspot.com, and he has coauthored a recent book called *Event Processing: Designing IT Systems for Agile Companies* (McGraw Hill, 2009). Chandy has a PhD in operations research and electrical engineering from the Massachusetts Institute of Technology. Contact him at mani@cs.caltech.edu.

The Play's the Thing

R. Michael Young
North Carolina State University

For most of the 20th century, our entertainment media — film, music, novels, and TV — were happily non-interactive. But a significant shift in the past 30 years toward interactive entertainment has built the computer game industry into a powerhouse that generates more than US$19 billion in annual revenue worldwide, rivaling both music sales and box office receipts. For most of this industry's history, games were primarily designed to be played alone, but even this has changed, with the single-player focus shifting in the past five years to exploit the increase in broadband availability and include additional players.

As computer and console games continue to exploit new services available via the Internet, the design of gameplay itself will correspondingly change. These changes will expand the already powerful social and cultural roles that games play as well as enable the development of new core game technologies involving 3D graphics, real-world/augmented reality interfaces, and artificial intelligence.

Playing in the Cloud(s)

From a market perspective, it's the players' desire for social connectivity that will drive the coming shift to networked gameplay. Already, developers of major game titles are marginalizing their single-player modes and focusing their development efforts on enhancing their networked multiplayer offerings. In fact, some high-profile games are now designed exclusively for online play. Although the shift toward network-enabled games is currently motivated by the desire to enhance gameplay with a social element, the added computational power the shift brings has much broader significance.

We can categorize the kinds of innovations we'll see in game development as a result of the increased access to network services as belonging to one of two types: those that make current high-end game capabilities available across a range of hardware and those that bring new

game capabilities into existence. In the former category, we're already seeing early steps to provide compute-intensive game services via the cloud — for instance, by shifting the graphics-rendering process from a player's PC to cloud-based render farms. In these approaches, a game's high-end 3D graphics are produced on remote servers and streamed as video to lightweight clients. In general, approaches like this will add new value to games by migrating conventional game computation from the player's machine to high-end servers, effectively raising the bar for compute power across all users. It will also allow, for instance, high-end virtual worlds, educational simulations, and serious games to run on low-end hardware in schools that lack modern computer laboratories and in the homes of families who can't afford today's high-end hardware.

Even more significantly, this shift to the cloud will provide access to compute services that will enable new types of intelligent tools to add value to games in ways we've only begun to explore. Exciting new techniques are currently being developed that let game engines create game content on-the-fly rather than requiring it to be crafted by hand and compiled into a game at design time. These methods, collectively referred to as *procedural content generation* (PCG), leverage computational models of in-game phenomena to generate content dynamically. Ongoing PCG research projects seek to build systems that can automatically create entire cities, forests full of diverse and unique trees and plants, and novel game character bodies that move smoothly according to anatomical and physiological constraints.

General methods for PCG are computationally costly and so have seen commercial use only in very limited contexts. By moving these functions to the cloud, PCG techniques bring this new functionality to the game client software at almost no cost. Furthermore, the use of cloud-based servers for PCG will promote the development of even more transformative uses of content generation, including complex character dialogue, dynamic 3D camera control, and complex and adaptive story generation. In the future, games that use PCG on remote servers will tailor each player's session to his or her preferences, goals, and context. Each city street you race down, each thug you interrogate, each quest your raiding party embarks on will be created on the spot to provide you with a carefully crafted entertainment experience.

Taking It to the Street

One of the most significant changes in interactive entertainment will arise from the combination of network-centric game services with powerful, pervasive, and location-aware handheld computing platforms and smart phones. This powerful combination will break down the boundary between play and many other aspects of our lives, making entertainment not just accessible during our leisure time but an integral part of our work, social life, shopping, and travel. Thanks to GPS, games running on mobile platforms will not only know who you are, but where you are, letting designers adjust a game's content and challenges to the physical/geographical space in which you're playing.

By relying on network services to manage a game's state, games will be designed to seamlessly slide from cell phone to game console to work PC to home media center as players move from context to context during the day. Social gameplay will be further enhanced by designing games that take into account other players located in the same physical space — for example, when riding on a city bus or touring a foreign city. Services that facilitate the easy creation of and access to location-specific data will make game content creators out of local governments, merchants, civic groups, and individuals. In the near future, your game will adapt to the political history of the village you're driving through, the goals of the anonymous player who's sharing your subway car, and the sale on khaki pants at the Gap that you just walked past.

The two network-centric aspects of games described here — the power of cloud computing and pervasive, location-aware connectivity — will change not just gameplay but will also alter the boundaries between entertainment and what we've traditionally thought of as more serious computing contexts. I expect to see a stronger integration of virtual spaces, information spaces, and real-world spaces. The pervasive nature of online interactive entertainment will push the interface metaphors and user experiences found in games into the broader context

of computing. It's clear that those broader contexts will change as a result. The challenge for game designers is to figure out how the broader contexts will, in turn, change games.

R. Michael Young is an associate professor of computer science and the co-director of the Digital Games Research Center at North Carolina State University. He has a PhD in intelligent systems from the University of Pittsburgh and is a member of the IEEE, the ACM, and the Association for the Advancement of Artificial Intelligence. Contact him at young@csc.ncsu.edu.

The Growing Interdependence of the Internet and Climate Change

Larry Smarr
University of California, San Diego

As proven by the global attendance at December's UN Climate Change Conference 2009 (http://en.cop15.dk/), more attention is being paid to the components of our society responsible for the emission of greenhouse gases (GHGs) and how to reduce those emissions. The global information and communication technology (ICT) industry, which includes the Internet, produces roughly 2 to 3 percent of global GHG emissions, according to the Climate Group's Smart2020 report (www.smart2020.org). Furthermore, if it continues to follow a business-as-usual scenario, the ICT sector's emissions will nearly triple by 2020.

However, the Climate Group estimates that the transformative application of ICT to electricity grids, logistic chains, intelligent transportation, building infrastructure, and dematerialization (telepresence) could reduce global GHG emissions by roughly 15 percent, five times ICT's own footprint! So, the key technical question before our community is, can we reduce the carbon intensity of Internet computing rapidly enough that even with its continued spread throughout the physical world, the ICT industry's overall emissions don't increase?

This is a system issue of great complexity, and to make progress we need numerous at-scale testbeds in which to quantify the many trade-offs in an integrated system. I believe our research university campuses themselves are the best testbeds, given that each is in essence a small city, with its own buildings, hospitals, transportation systems, electrical power generation and transmission facilities, and populations in the tens of thousands. Indeed, once countries pass legislation for carbon taxes or "cap and trade" markets, universities will have to measure and reduce their own carbon footprints anyway,[1] so why not instrument them now and use the results as an early indicator of the optimal choices for society at large?

As discipline after discipline transitions from analog to digital, we'll soon find that when the carbon accounting is done, a substantial fraction of a campus's carbon footprint is in its Internet computing infrastructure. For instance, a major carbon source is data center electrification and cooling. Many industries, government labs, and academics are working to make data centers more efficient (see http://svlg.net/campaigns/datacenter/docs/DCEFR_report.pdf). At the University of California, San Diego (UCSD), our US National Science Foundation-funded GreenLight project (http://greenlight.calit2.net) carries this work one step further by providing the end user with his or her application's energy usage. We do this by creating an instrumented data center that allows for detailed real-time data measurements of critical subcomponents and then making that data publically available on the Web, so that the results can guide users who wish to lower their energy costs.

This is more complex than you might think at first. Any given application, such as bioinformatics, computational fluid dynamics, or molecular dynamics, can be represented by several algorithms, each of which could be implemented in turn on a variety of computer architectures (multicore, field-programmable gate array, GPUs, and so on). Each of these choices in the decision tree requires a different amount of energy to compute. In addition, as UCSD's Tajana Rosing has shown, we can use machine learning to implement various power[2] or thermal[3] management approaches, each of which can save up to 70 percent of the energy used otherwise in the computations.

Another strategy to reduce overall campus carbon emissions is to consolidate the clusters and storage systems scattered around campus in different departments into a single energy-efficient facility and then use virtualization to increase the centralized cluster's utilization.

We could also use zero-carbon energy sources (solar or fuel cells), which produce DC electricity, to drive the cluster complex, bypassing the DC to AC to DC conversion process and reducing the operational carbon footprint of campus computing and storage to zero.

As we reduce the carbon emissions required to run Internet computing, we can extend the Internet into new functions, such as instrumenting buildings for their energy use and eventually autonomously controlling building systems in real time to reduce overall energy use. An example is the research performed in UCSD's Computer Science and Engineering building by Rajesh Gupta and his colleagues, who found that roughly 35 percent of the building's peak electrical load is caused by PCs and servers. His team's research also showed that intelligent sleep-state management could help avoid a large fraction of this Internet computing electrical load (www.usenix.org/events/nsdi09/tech/full_papers/agarwal/agarwal_html/).

Another application of Internet computing to avoid carbon emissions is *dematerialization*, such as using Internet video streaming to reduce air or car travel to meetings. At Calit2, we use a variety of compressed high-definition (HD) commercial systems such as LifeSize H.323 videoconferencing (approximately 1 to 2 Mbps) or high-end systems such as Cisco's Telepresence system (approximately 15 Mbps). However, we're also experimenting with uncompressed (1,500 Mbps) HD (developed by the University of Washington's Research Channel) or with digital cinema (four times the resolution of HD), which requires 7,600 Mbps uncompressed! These higher-bandwidth video streams are used over dedicated optical networks (such as CENIC, Pacific Wave, the National LambdaRail, Internet2's Dynamic Circuits, or the Global Lambda Integrated Facility, all operating at 10,000 Mbps).

We can extend the notion of virtual/physical spaces from simple face-to-face meetings to creating collaborative data-intensive analysis environments in which whole rooms are "sewn together" using the Internet video streaming technologies mentioned earlier. Calit2 is an institute that spans two University of California campuses, San Diego and Irvine, separated by a 90-minute drive. We recently started using HD streaming video to link our two auditoriums together for joint meetings, such as our all-hands meetings. Previously, we needed dozens of people from one campus to drive to the other campus for such a meeting.

Another example that focuses more on research is how Calit2 in San Diego and the NASA Ames Lunar Science Institute in Mountain View, California, have both set up large tiled walls (displaying tens to hundreds of megapixels) called OptIPortals and then used the CENIC dedicated 10-Gbps optical networks to couple their two rooms with streaming video and spatialized audio. This lets researchers at both ends explore complex lunar and Martian images taken by orbiting or surface robotic craft. Each side can control image placement and scaling on the other's wall, so team brainstorming is as easy as if both sides were in the same physical room. We use this on a weekly basis, avoiding a significant amount of plane travel and the carbon emissions that it would otherwise produce.

The world must act this coming decade to make drastic changes in the old "high carbon" way of doing things and transition to a new "low carbon" society.

These ideas are just the tip of the iceberg of how we can turn our research universities into living laboratories of the greener future. As more universities worldwide begin to publish their results on the Web, best practices will quickly develop and lessons learned can be applied to society at large. This is essential because the world must act this coming decade to make drastic changes in the old "high carbon" way of doing things and transition to a new "low carbon" society if we're to avoid ever worsening global climatic disruption.

References

1. B. St. Arnaud et al., "Campuses as Living Laboratories for the Greener Future," *EDUCAUSE Rev.*, vol. 44, 2009, pp. 14–32.
2. G. Dhiman and T. Simunic Rosing, "System-Level Power Management Using Online Learning," *IEEE*

Trans. Computer-Aided Design, vol. 28, no. 5, 2009, pp. 676–689.

3. A. Coskun and T. Simunic Rosing, "Proactive Thermal Management," to be published in *IEEE Trans. Computer-Aided Design*, 2009.

Larry Smarr is the Harry E. Gruber Professor in the Department of Computer Science and Engineering in the Jacobs School of Engineering at the University of California, San Diego, and director of the California Institute for Telecommunications and Information Technology, a UC San Diego/UC Irvine partnership. His research interests include green IT, global telepresence, and microbial metagenomics. Smarr has a PhD in physics from the University of Texas at Austin. Contact him at lsmarr@ucsd.edu or follow him as lsmarr on Twitter.

The Internet and Past and Future Communications Revolutions

Andrew Odlyzko
University of Minnesota

In attempting to predict the Internet's evolution in the next decade, it is instructive, as well as amusing, to consider the previous January/February 2000 Millennial Forecast issue of *IC*. The authors were all Internet luminaries with sterling records. Yet, although there were many perceptive and accurate comments in their essays, most of their predictions turned out to significantly miss the mark. In many cases, this came from overestimates of the speed of change, a tendency that's almost universal among inventors and promoters of new technologies. As just one example, Bill Gates predicted that books would "go digital ... broadly in the next five years." With the arrival of the Amazon Kindle and other e-readers, we're probably finally seeing the start of this transformation, but it now seems safe to say that a broad move toward digital books is at least five years further out, 15 years after Gates made his original forecast. Many other predictions seem in retrospect to have been completely misguided. For example, Eric Schmidt, at the time head of Novell, touted a secure worldwide "distributed directory service" as the "master technology" of the next wave on the Internet. Yet such a service is nowhere in sight, and Schmidt at his current position at Google has found how to gain profit and influence through insecure but workable statistical approaches to serving the needs of the public and advertisers.

The lack of accuracy in the previous forecast issue shouldn't be a surprise. History is replete with examples of the difficulty of forecasting how quickly technologies will advance and how society will use them. What is less well known, though, is how oblivious people can be to the massive moves taking place around them that affect their industries.

The previous forecast issue had just one discussion of voice, in Jim White's essay, which predicted that voice browsers would become widespread and important. Yet, the big communications revolution that was taking place then, and has continued over the past decade, overshadowing the Internet the entire time, has been the growth in mobile voice. In the US alone, wireless industry revenues reached almost $150 billion in 2008 — that's roughly four times the revenue from US residential high-speed Internet access. It's also almost twice the worldwide revenue that Hollywood's entertainment offerings enjoyed and almost seven times Google's worldwide revenues that year. What the market clearly shows is that narrowband mobile voice is far more important to people, in terms of either the number of users or their willingness to pay, than broadband Internet access. (The figures for the rest of the world, especially the less developed countries, are skewed even more dramatically in favor of wireless voice over Internet, both in users and in revenue.)

Advances in mobile technologies are providing higher transmission capacities and are leading to a convergence of wireless with IP, the Internet Protocol. This has obvious implications for wireless, but potentially even more important, if less obvious, implications for the Internet. Of the 20 percent of wireless revenues that don't come from voice, the lion's share is from texting, which, just like voice, is a simple connectivity service. This shows that you can be successful simply by providing dumb pipes. However, wireless carriers misunderstand this, and claim their success is due to the tight control they exercise over their networks. They worry that bringing Internet technologies to their industry will lead to their business becoming a commodity-based one and are trying to

tightly control just what services are offered over their channels. The extent to which they succeed will be fascinating to watch but seems impossible to predict because it depends far less on technology than on economics, industry structure, and regulation.

Many of the old battles over issues such as quality of service, flat rate versus usage-based pricing, and the ongoing war over net neutrality are likely to be fought again in the wireless arena, and it's possible that some outcomes might be different this time. The reason is that the balance between supply and demand is different: in the wireline arena, the growth in demand is still high, but it has been declining to a level that's currently just about counterbalanced by technological improvements. This produces incentives for service providers to increase usage, and such incentives suggest simple pricing and simple networks.

In wireless, though, growth in data transmission appears to be significantly ahead of what technology can support, at least without major increases in capital expenditure. The incentives to raise such investments are lacking because most of the large potential sources of new wireless data transmissions aren't anywhere near as lucrative as voice and texting. Users want seamless mobility, but the huge gap between the capacities of fiber and radio links is unlikely to allow it, so service providers have strong incentives to closely manage their network traffic and are likely to try to ration capacity. Network management will be especially important to protect the cash cow — voice.

On the other hand, many of the incentives toward open networks that have so far prevailed in the wireline Internet do apply, and will continue to apply, in mobile data. Service providers and their system suppliers have demonstrated repeatedly that they're terrible at service innovation. They have neglected numerous opportunities, even in basic services — for example, by not providing higher-quality voice. And their major successes, such as texting and ring-tone downloads, were accidents, not something they planned.

The AT&T deal with Apple over the iPhone could be a sign that the wireless industry is beginning to acknowledge its limitations and is willing to open a door to more innovative outsiders. But the battles for control surely won't go away. (Even the iPhone deal involves considerable control, by Apple this time. It isn't a fully open ecosystem.) For the wireline Internet, the convergence of IP with wireless could have various unanticipated outcomes. Because mobility has great value for users, spending might tilt even further in that direction, and consequently, innovation could shift to the wireless arena (whether in an open environment or in a collection of walled gardens). New services might be designed primarily for the relatively low-bandwidth wireless sector, not for the big pipes available in wireline, which might end up as a backwater. That said, the availability of wireless Internet access could spur growth of very high-bandwidth wireline access even in countries currently lagging in that field. Some wireline providers, especially those with dominant wireless operations, might stop upgrading their land lines, but others could find that the only way to survive is to exploit their one natural advantage: ability to provide big pipes with low latency.

The battles for control surely won't go away. For the wireline Internet, the convergence of IP with wireless could have various unanticipated outcomes.

The only safe bet is that service providers will continue repeating many of their old mistakes — in particular, their preoccupation with content as opposed to connectivity. But beyond that, predictions appear even harder to make than a decade ago, as there are more options before us.

Andrew Odlyzko is a professor in the School of Mathematics at the University of Minnesota. His research interests include number theory, cryptography, and economics of data networks. Odlyzko has a PhD in mathematics from MIT. Contact him at odlyzko@umn.edu or via www.dtc.umn.edu/~odlyzko.

See IC*'s millennium predictions in the January/February 2000 special issue: http://doi.ieeecomputersociety.org/10.1109/MIC.2000.815848.* — Ed.

Fighting over the Future of the Internet

David Clark
MIT Computer Science and Artificial Intelligence Laboratory

For much of the Internet's life, it has coevolved with the PC. The relative maturity of the PC could thus lead to the erroneous assumption that the Internet itself is mature. But as computing enters the post-PC era over the next decade, with mobile devices, sensors, actuators, and embedded processing everywhere, the Internet will undergo a period of rapid change to support these new classes of computing.

Even as the evolving nature of computing changes the Internet, the more important drivers of change are likely to be economic, social, and cultural. I have written previously about how a set of tussles among various stakeholders will define the Internet's future.[1] Going forward, what might those tussles be and what might they mean for the Internet?

As we predict the future, we should not underestimate the importance of cultural issues (and cultural differences in particular). Technical systems based on technical standards usually work the same everywhere. Such homogeneity can directly collide with divergent norms about such things as openness, privacy, identity, intellectual property protection, and, perhaps most fundamentally, the balance of rights between the state and the individual. One possible outcome is the acceptance of Internet hegemony as a force for cultural uniformity. But, especially because that force is sometimes equated with the unwelcome cultural and political hegemony of the US — such things as our entertainment industry, our sports icons, and our fast food (not to mention our language and politics) — you can see the potential for a backlash that leads to a fragmentation of the Internet into regions, where behavior within regions is consistent with each region's norms, and connectivity is more constrained among regions. The drivers of this backlash would be nation-states trying to preserve their sovereignty and jurisdictional coherence, aligned with a grassroots desire to preserve cultural diversity. In a world where nations seem to fight wars as much over identity as economics, the alignment of these forces can have significant consequences.

Issues of economics and industry structure will also drive change. To model the future, it is helpful to catalog the tussles of today. There is a fundamental tussle between a core value of the current Internet — its open platform quality — and investors' desire to capitalize on their investments in expensive infrastructure. Examples include debates over network neutrality, debates over whether ISPs and their business partners can profile their customers for advertising purposes, and the collision between the open Internet and more closed sorts of networks such as those for video delivery.

Looking forward, we must steer between two perils to reach a healthy and vibrant future. If ISPs, in pursuit of additional revenues, diverge from the Internet tradition of the open neutral platform and favor their preferred content and applications over those of unaffiliated third parties, it might reduce the rate of innovation, reduce the supply of content and applications, and stall the Internet's overall growth. On the other hand, if (perhaps due to regulation) ISPs provide only "simple dumb pipes," a commodity business fraught with low margins, they might not see a reason to upgrade their facilities, which could lead to stagnation in the capabilities or scale of the Internet. Both outcomes are unwelcome and feared by different stakeholders. In my view, there is a middle path that avoids both perils, but we will have to steer carefully down that path, especially if we are going to impose explicit regulation.

It is important to remember that the shape of tomorrow's ISP is not fixed and mature, any more than the Internet itself. Different business and policy decisions will have major influences on the future.

The packet layer of the Internet is not the only platform over which we will tussle. In the future, we might well debate whether higher-level application development platforms such as the iPhone or Facebook should be limited in the control they can impose. The Internet, taken broadly rather than just as a packet mover, is layers of platform on platform. One prediction about the future is that the debates will move "up" in the layers.

The research community today is exploring new concepts for networking, based on alternative modes of basic interaction — for example, delay-tolerant networks (DTNs), which relay data in a series of stages rather than directly from origin to destination, and

information dissemination networks and "publish/subscribe" paradigms, which have a similar "staged" character. These modes will prove very important in the future Internet, and we will fight over whether they are competitive offerings "on top of" the basic packet forwarding paradigm or whether they become the basic service paradigm, in which case the owners of the network have the exclusive ability to provide them. I personally believe that we will be much better off if application designers (and users) can select from a suite of competing service offerings at these levels. But this is just another example of the tussle over platform ownership.

To add another dimension to all these tussles, consider the future of the developing world. We have different governments with different cultures and rules of regulation, different users with different skills, using perhaps different platforms (such as mobile devices) with different histories of open access and business models for investment. The result is a rich and heterogeneous stew of expectations, onto which we will try to impose uniform Internet standards.

I mentioned that the future Internet will link not just PCs and PDAs but also sensors and actuators. An important tussle that we must anticipate is the one associated with the phrase "the surveillance society." Networked sensors have the ability to change society in fundamental ways, but those changes and the tensions they raise will have the power to change the Internet. At a technical level, the volume of data from sensors (including video monitors) could swamp the current sources of data today — business practice and human endeavor. At a policy level, one could imagine that the ISPs in one or another country are assigned to control access to various sorts of data (sensor and other), or alternatively, authoritatively add certain sorts of metadata to data from sensors. We must expect tensions over the embedding (or not) of data about geolocation, identity, information authenticity, access rights or limits, and so on into one or another protocol.

While I called the tussle over open platforms fundamental, the tussle between the state and the individual is even more fundamental. The Internet has been glorified as a tool for personal empowerment, the decentralization of everything, collective action, and the like. It has empowered nongovernmental organizations and transnational actors. But at the same time, IT is a tool for information gathering and processing, and modern government is essentially a data-driven bureaucracy. In the hands of government, the Internet and, more generally, the tools of cyberspace are a force for centralized control. So, we see the dual ideas of decentralization and user empowerment on the one hand doing battle with images of the surveillance society and total information awareness on the other. Are the individuals (and those who advocate for them) powerful enough to resist the tensions from the center? That is a critical question in scoping the future.

Finally (and perhaps beyond the time frame of these essays), we should ask what is next after the Internet has hooked up all the sensors, all the actuators, all the cars, and all the smart dust. The next step must be hooking up humans. As we have more and more computers about us, and then inside us (as will certainly happen), all those computers will benefit from networking. This takes us to the debates over human augmentation, which will bring ethical and religious elements into the tussle. The future is going to be exciting.

In the hands of government, the Internet and, more generally, the tools of cyberspace are a force for centralized control.

Reference

1. D.D. Clark et al., "Tussle in Cyberspace: Defining Tomorrow's Internet," *Trans. IEEE/ACM Networking*, vol. 13, no. 3, 2005, pp. 462–475.

David Clark is a senior research scientist at the MIT Computer Science and Artificial Intelligence Laboratory. His research interests include the future of the Internet, network security, and the implications of economics and society on technology. Clark has a PhD in computer science from MIT. Contact him at ddc@csail.mit.edu.

Back to the Future of Internet

Viviane Reding
European Commission

As we brace ourselves for the Internet's future, policy challenges are building up at a staggering speed. We must act now so that in 10 years' time, we don't wake up one morning and realize that we would need to go back to 2010 to fix everything. If, in the future, the Internet can't cope with the amount of data we want to send, we aren't able to use it in the nomadic way we prefer today, or our privacy isn't protected, the Internet won't reach its full potential to improve daily life and boost the world economy. If this happens, it will be due to serious policy failures, not the technology itself. Citizens will rightly ask policy makers and the Internet community, "What went wrong?"

My vision for the Internet's future: it should be open to all sorts of technological innovation; it should be fast enough to allow all sorts of new uses; it should be reliable and secure so that important services can be carried out online; it should be available to everyone; and it should be a place where everyone can express their ideas freely.

The Phoenix

Europe has set itself the challenge of promoting growth while ensuring a smooth transition to a low-carbon resource-efficient economy. This challenge is also a great opportunity, especially for technology sectors such as telecoms and the Internet. High-speed broadband Internet offers a great chance for smart investment that will help a speedy recovery from recession but also make economies more competitive for the next decade. New resource-efficient growth based on high-speed Internet can rise from the smoldering ashes of the current credit crunch.

In 2013, the Internet will be nearly four times larger than it is today.[1] Applications such as remote data backup, cloud computing, and video conferencing demand high-speed access. Cloud computing will unleash the potential of small- and medium-sized enterprises (SMEs) worldwide. Europe, which is heavily dominated by SMEs as compared to the US, will surely benefit when companies can rent, rather than buy, IT services. A recent study[2] estimated that such online business services could add 0.2 percent to the EU's annual gross domestic product growth, create a million new jobs, and allow hundreds of thousands of new SMEs to take off in Europe over the next five years. The Internet also offers a platform for a new wave of smart and green growth and for tackling the challenges of an aging society.

Green and Sustainable Growth

Information and communication technology (ICT) offers a significant toolset to build a sustainable future for the next generation of Europeans. To meet its ambitious climate change goals, the EU strongly endorses new technologies capable of improving energy efficiency and making transport systems, buildings, and cities in general smarter.

Smart Energy Grids

Electricity generation around the world is expected to nearly double in the next 20 years, from roughly 17.3 trillion kilowatt-hours (kWh) in 2005 to 33.3 trillion kWh in 2030 (www.eia.doe.gov/oiaf/ieo/pdf/electricity.pdf). Today, up to 40 percent of the energy produced might be lost on its way to the consumer, but Internet connectivity, computing power, digital sensors, and remote control of the transmission and distribution system will help make grids smarter, greener, and more efficient. These smart grids can also integrate new sources of renewable power, allow coordinated charging of devices, and give consumers information about how much energy they use. In turn, this helps energy companies control their networks more effectively and reduce greenhouse gas emissions. Some pilot projects, using today's Internet technologies, have already reduced peak loads by more than 15 percent — imagine what would happen with tomorrow's technology (www.oe.energy.gov/DocumentsandMedia/DOE_SG_Book_Single_Pages(1).pdf).

Smart Transport Systems

Traffic jams cost Europe €135 billion a year, and drivers lose five days per year while sitting in traffic. Simply building new roads isn't the solution — making roads and cars "smarter" with intelligent transport systems (ITS) such as sensor networks, RF tags, and positioning systems offers a promising alternative. The Internet can interconnect diverse technologies and make mobility more efficient through the

real-time management of public and private transport resources, traveler information, and decision-making tools.

Smart Healthcare Systems

Current research is working to develop technologies for "ambient" environments capable of assisting patients by treating and monitoring them from a distance to reduce medical costs and improve patient comfort. These technologies combine devices (sensors, actuators, special hardware, and equipment), networks, and service platforms to harness information about medical conditions, patient records, allergies, and illnesses.

Is the Internet Broken?

The Internet was never really designed to meet the variety of demands we place on it. Many experts recognize that it's almost stretched to a breaking point — in particular, because of soaring amounts of content and traffic and new demands for mobile Internet access. New usage patterns and markets are generating Internet bottlenecks and demands for infrastructure and technology upgrades. However, there's very little incentive for telecom operators to invest in infrastructure, and business models — especially for Web content — are uncertain.

European industry leaders and scientists have made Web 3.0 Internet services a top research priority. Future Internet technologies — the key to Web 3.0's success — are a core focus of the EU's overall research program (Framework Programme 7; http://cordis.europa.eu/fp7/ict/programme/); the European Commission's FIRE (Future Internet Research and Experimentation; http://cordis.europa.eu/fp7/ict/fire/home_en.html) initiative also encourages long-term investigation and experimental validation of new and visionary Internet concepts.

The Internet of Things, which interconnects objects from books to cars to electrical appliances to food, will also require a radical rethink of how the Internet operates. To create the space needed to address the coming explosion of connected devices, it's imperative that we make the transition to IPv6 and avoid compromising the Internet's ability to securely offer more functions and services. Policy makers must now call on industry to make this transition and make sure that public administrations migrate to IPv6.[3]

It's difficult to see alternatives to the private sector's traditional dominance over the Internet's management. This system has supported us well, but we must work to let governments worldwide exercise their responsibilities by balancing the US's national priorities with the international community's legitimate expectations and interests.

Who Do We Trust?

With ICT services becoming more pervasive and not only extending into new areas of our private lives but also being an indispensable tool for the daily running of business operations, privacy, security, and empowerment are now imperative. Moreover, getting digital trust and confidence right could provide an additional 11 percent growth (or €46 billion) on top of the natural expected growth of the EU's digital economy. Failure to act, however, can result in greater damages — 18 percent of growth (€78 billion) could be lost or significantly delayed.

The US and the EU have somewhat different approaches to protecting privacy, but I'm convinced all Internet users would prefer to be empowered to make decisions concerning their own privacy. The European Commission recently published a recommendation on protecting privacy on the Internet of Things (http://europa.eu/rapid/pressReleasesAction.do?reference=IP/09/740). By default, it mandates that a smart tag (RFID device) attached to an object that a consumer is about to buy is disabled at the point of sale, unless the consumer explicitly agrees to keep the tag active. (There are some benefits to leaving RFID tags active — for instance, telling a washing machine how to wash your sweater.) The worldwide market value for RFID tags is estimated to have been €4 billion in 2008 and is predicted to grow to roughly €20 billion by 2018.[4] However, it's only via consumer acceptance that this market's potential can be realized, and this acceptance demands trusted services.

Empowering users is not only paramount for privacy — it's also essential when organizations design new business models based on converging services. Consumer choice shouldn't be artificially limited; instead, let the consumer decide and determine innovation.

Words Are No Substitute for Action

Openness is one of the main ingredients of an innovative Internet: this key characteristic

shouldn't be compromised because of the architecture's future evolution. We must take advantage of open interfaces and standards, so that markets can grow without forcing consumers to use a certain kind of software or application or pay unnecessary royalties.

Net neutrality is also essential. The debate in the US has been extensive, and to an extent is mirrored in Europe's discussions about telecom rules. When new network management techniques allow traffic prioritization, we must prepare for these tools being used for anticompetitive practices. With the new telecom regulatory framework, approved in November 2009, the European Commission empowered national regulators to prevent such unfair abuse to the detriment of consumers. It's a good first step, but we must keep monitoring this issue as we progress.

We need only look at recent events in Myanmar and Iran to see how much the Internet has become an important vehicle for sharing political views, even be those with minority, controversial, or even censored opinions. The Internet's instantaneous nature allows users to post event information or eye-witness accounts in almost real time. Naturally, this poses challenges for those entities trying to restrict access to information that isn't convenient for them. For its part, the European Commission is determined to promote freedom of speech wherever possible in its international relations.

Many of these issues are fundamentally of a global nature and deserve a global discussion. If we design and prepare for the future Internet, putting the user at the center and keeping it open, we won't have to back track and fix anything because of the decisions we made in 2010. Going back to the future isn't only impossible, it's also a sign of poor global ICT leadership.

References

1. "Cisco Visual Networking Index: 2008-2013," Cisco, June 2009; www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html.
2. F. Etro, "The Economic Impact of Cloud Computing on Business Creation, Employment and Output in Europe," *Rev. Business and Economics*, 2009, p. 3.
3. "Communication from the Commission – Action Plan for the Deployment of Internet Protocol," *IPv6 in Europe – COM*, vol. 313, 2008, pp. 8–9.
4. Booz and Company, "Digital Confidence, Securing the Next Wave of Digital Growth," *Liberty Global Policy Series*, 2008, p. 4.

Viviane Reding is the European Commissioner for Information Society and Media. She has a PhD in human sciences from the Sorbonne (Paris). Contact her at viviane.reding@ec.europa.eu.

Intercultural Collaboration Using Machine Translation

Toru Ishida
Kyoto University

Almost every country on Earth is engaged in some form of economic globalization, which has led to an increased need to work simultaneously in multiple cultures and a related rise in multilingual collaboration. In local communities, we can already see this trend emerging in the rising number of foreign students attending schools. Regional communities have had to solve the communication problems among teaching staffs, foreign students, and their parents, typically by focusing on relieving culture shock and its related stress with the aid of bilingual assistants. When turning our eyes to global communities, problems such as the environment, energy, population, and food require something more – mutual understanding. In both local and global cases, the ability to share information is the basis of consensus, thus language can be a barrier to intercultural collaboration.

Because there's no simple way to solve this problem, we must combine several different approaches. Teaching English to both foreign and local students is one solution in schools, but learning other languages and respecting other cultures are almost equally important. Because nobody can master all the world's languages, machine translation is a practical interim solution. Although we can't expect perfect translations, such systems can be useful when customized to suit the communities involved. To customize machine translations, however, we need to combine domain-specific and community-specific dictionaries, parallel

texts with machine translators. Furthermore, to analyze input sentences to be translated, we need morphological analyzers; training machine translators with parallel texts requires dependency parsers. In the future, users might also want to use speech recognition/synthesis and gesture recognition. Even for supporting local schools, which include students from different countries, we need worldwide collaboration to generate all the necessary language services (data and software). Fortunately, Web service technologies enable us to create a workflow that assists in their creation. At Kyoto University and NICT, we've been working on the Language Grid,[1] which is an example of a service-oriented language infrastructure on the Internet.

Customized Language Environment Everywhere

Let's look at what could happen in the very near future in a typical Japanese school, where the number of Brazilian, Chinese, and Korean students is rapidly increasing. Suppose the teacher says "you have cleanup duty today (あなたは今日掃除当番です)" in Japanese, meaning "it is your turn to clean the classroom today." Now imagine that some of the foreign students don't understand what she said — to figure it out, they might go to a language-barrier-free room, sit in front of a computer connected to the Internet, and watch the instructor there type the following words in Japanese on the screen: "you have cleanup duty today." The resulting translation appears as "今天是你负责打扫卫生" in Chinese, "오늘은 네가 청소 당번이야" in Korean, and "Hoje é seu plantão de limpeza" in Portuguese. "Aha!" say the kids with excited faces. One of them types in Korean, "I got it," and the translation appears in Japanese on the screen.

Is machine translation that simple to use? Several portal sites already offer some basic services, so let's challenge them with my example from the previous paragraph. Go to your favorite Web-based translation site and enter, "you have cleanup duty today" in Japanese and translate it into Korean. But let's say you're a Japanese teacher who doesn't understand Korean, so you aren't sure if the translation is correct; to test it, you might use back translation, clicking on the tabs to translate the Korean translation back into Japanese again, which yields, "you should clean the classroom today." It seems a little rude, but it might be acceptable if accompanied with a smile. Let's try translating the Chinese translation in the same way. When we back translate it into Japanese, we might get the very strange sentence, "today, you remove something to do your duty." It seems the Japanese word "cleanup duty" isn't registered in this machine translator's dictionary.

Basically, machine translators are half-products. The obvious first step is to combine a domain-specific and community-specific multilingual dictionary with machine translators. Machine-translation-mediated communication might work better in high-context multicultural communities, such as an NPO/NGO working for particular international issues. Computer scientists can help overcome language barriers by creating machine translators that generalize various language phenomena; multicultural communities can then customize and use those translators to fit their own context by composing various language services worldwide.

Issues with Machine-Translation-Mediated Communication

Even if we can create a customized language environment, we still have a problem in that most available machine translators are for English and some other language. When we need to translate Asian phrases into European languages, we must first translate them into English, then the other European language. If we use back translation to check the translation's quality, we must perform translation four times: Asian to English, English to European, and back to English and then to the original Asian language. Good translation depends on luck — for example, when we translate the Japanese word "タコ," which means octopus, into German, the back translation returns "イカ," which means squid, two totally different sushi ingredients.

The main reason for mistranslation is the lack of consistency among forward/backward translations. Different machine translators are likely to have been developed by different companies or research institutions, so they independently select words in each translation. The same problem appears in machine-translation-mediated conversation: when we reply to what a friend said, he or she might receive our words as totally different from what we actually, literally said. *Echoing*, an important tool for the ratification process in lexical entrainment (the

process of agreeing on a perspective on a referent) is disrupted, and it makes it difficult to create a common ground for conversation.[2]

Even if translation quality increases, we can't solve all communication problems through translation, so we must deepen our knowledge of different cultures to reach an assured mutual understanding. For example, we can translate the Japanese term "cleanup duty" into Portuguese, but it can still puzzle students because there's no such concept in Brazil. As is well known, deep linkage of one language to another is the first step in understanding, thus we need a system that associates machine translation results with various interpretations of concepts to help us better understand different cultures. I predict that Wikipedia in particular will become a great resource for intercultural collaboration when combined with machine translators because a large portion of Wikipedia articles will be provided in different languages and linked together.

References

1. T. Ishida, "Language Grid: An Infrastructure for Intercultural Collaboration," *IEEE/IPSJ Symp. Applications and the Internet* (SAINT 06), IEEE Press, 2006, pp. 96–100.
2. N. Yamashita et al., "Difficulties in Establishing Common Ground in Multiparty Groups using Machine Translation," *Proc. Int'l Conf. Human Factors in Computing Systems* (CHI 09), ACM Press, 2009, pp. 679–688.

Toru Ishida is a professor in Kyoto University's Department of Social Informatics and a leader of the NICT Language Grid project. He has been named an IEEE fellow for his contribution to autonomous agents and multiagent systems. His research interests include social information systems, digital cities, and intercultural collaboration. Ishida has a PhD in computer science from Kyoto University. Contact him at ishida@i.kyoto-u.ac.jp.

The New Way of Business

Sharad Sharma
Canaan Partners

Every 20 years or so, a set of enabling technologies arrives that sets off a major organizational change. For example, new PCs, cheap networking, and WANs enabled organizations to adopt new processes and both decentralize and globalize much more easily than at any other point in the past. If we look ahead, in the next 10 years, Internet, mobile, and cloud computing will conspire to change things around us even more. New applications will emerge, which in turn will drive changes in how business is conducted.

Bookends of Change

We're reaching a significant tipping point from a connectivity viewpoint. Today, a little more than 2 billion people connect with each other over the Internet or mobile phones; by sometime in 2011, we can expect another billion, including farmers in India and fishermen in Indonesia (http://communities-dominate.blogs.com/brands/2009/03/the-size-of-the-mobile-industry-in-2009-short-overview-of-major-stats.html). By some estimates, we could have 5 billion people connected to the network by 2015, which will turn business and marketing models on their heads (www.connect-world.co.uk/articles/recent_article.php?oid=Global_2008_08).

Microfinance is already revolutionizing the provision of financial services to low-income clients and the self-employed, who traditionally lack access to banking and related services. One of the leading players in this space — and the first to go public — is SKS Microfinance, based in India. David Schappell of Unitus, a nonprofit microfinance accelerator, likens SKS to the small coffee shop that became Starbucks (www.time.com/time/magazine/article/0,9171,1186828,00.html). Not surprisingly, SKS is quite technology savvy: it uses smart cards, cheap mobile phones, and sophisticated back-end software to scale its business.

SKS represents the new breed of startups. They believe in the power of technology and of markets to bring about a large-scale, catalytic impact to the bottom of the pyramid. Their innovations in creating efficient marketplaces and new delivery systems for education, healthcare, and government services will blowback into more mainstream markets in the next 10 years.

At the other end of the spectrum, cloud computing is accelerating the quant revolution and

taking it to newer areas. The breadth of impact is staggering, ranging from in silico drug discovery to evidence-based medicine to experiment-based ad hoc creative design to social media analytics-based public relations (PR) to personalized recommendations in e-tailing, among others. The application of mathematics to problems of business and academia has been increasing rapidly over the past decade. Industries such as advertising and PR are in the throes of a transformation; social sciences are becoming data driven, and even philanthropy is embracing metrics.

Naturally, this has led to a tsunami of data. Increasingly, every part of the company organism is connected to the data center, and so every action — sales leads, shipping updates, support calls — must be stored. On top of this, the Web is surging with data, as machines from servers to cell phones to GPS-enabled cars manufacture updates. Fortunately, developing actionable insights from dynamic, dense data has become easier, partly because of the emergence of open source programming languages and tools. Moreover, the entry costs for computational infrastructure have come down due to cloud computing service providers.

Michael Lewis' book, *Moneyball: The Art of Winning an Unfair Game* (W.W. Norton & Co., 2003), turned many managers into converts of the quantitative method. Wall Street has offered another example: although some financial innovations turned out to be undesirable, it's difficult to deny the power of this intellectual revolution.

Inflexions that Matter

The rise of Internet, mobile, and cloud computing will bring lots of new micro-consumers into the fold and will reinvigorate innovation in mature industries. This presages some fundamental changes in how business will be conducted in the future. As Paul Saffo, a technology forecaster and consulting associate professor at Stanford University says, inflexion points are tip toeing past us all the time (http://blog.longnow.org/2008/01/14/paul-saffo-embracing-uncertainty-the-secret-to-effective-forecasting/). Based on what I've seen so far, I have some predictions for the next 10 years.

The Firm as Managed Network

Traditionally, the firm has been a collection of cost centers, but the pervasiveness of business metrics at the activity and function levels within the firm have made a different approach possible. In this approach, the firm can organize itself as a collection, indeed as a network, of profit centers. Because each profit center now has a clear financial incentive to drive higher productivity, tremendous efficiency is unleashed, translating into cost leadership for the firm. The poster child for this approach is Bharti Airtel, the largest mobile operator in India, which has 110 million subscribers. It makes a healthy 38 percent margin on average revenue per user of only US$7. In many ways, it has brought the same change that Southwest Airlines set off in the airline industry. As firms reconstitute themselves into managed networks of profit centers to become leaner, we will witness cost-structure transformations in many more industries.

Rise of Micro-Multinationals

Since the early 1990s, we've come a long way in the globalization of innovation. Today, the high-tech industry is extremely global: Nokia's biggest competitors are Asian (Samsung) and American (Apple); SAP and Oracle are at each other's throats but based in different continents. Global innovation is happening, and it's leading to global competition. To leverage this distribution innovation, firms of every size will organize themselves as multinationals.

Knowledge Jobs 2.0

Today's knowledge workers must interact with other companies, customers, and suppliers. Although some of these interactions are routine or transactional in nature, a growing number of them are complex, and complex interactions typically require people to deal with ambiguity and exercise high levels of judgment. These new knowledge workers will have to draw on deep experience, what economists call *tacit knowledge*. Jobs that include tacit interactions as an essential component are growing two-and-a-half times faster than the number of transactional jobs. In this interaction economy, vertically oriented organizational structures, retrofitted with ad hoc and matrix overlays, will be ill suited to modern work processes. So the traditional command and control paradigm is on its way out. A new management paradigm of connect and collaborate will take its place.

Creation Nets Will Multiply

Linux, Wikipedia, and Firefox show the power of self-organizing open communities. These networks of creators — in which thousands of participants come together to collaborate to create new knowledge — appropriate and build on each other's work. These creation nets are neither new nor limited to the software industry; their roots go back as far as the Italian Renaissance, when networks of apparel businesses in Piedmont and Tuscany sparked rapid innovation in the techniques of producing silk and cotton fabric. Today, an apparel network created by Li and Fung in China has 7,500 producers specializing in different areas that collaborate with each other. All these networks are predicated on rules of sharing that the participants established among themselves, a task that requires their mutual trust. Earlier, this trust building was possible only among people who lived in the same place, such as in Tuscany, or were affiliated with a credible network organizer. Now, the rules can be established among a much more diverse set of participants because of new communication technologies. Expect this to drive even more growth in creation nets.

These are just some of the seismic changes afoot that will make the next 10 years very exciting. In some ways, the die is set and the future is already here. It just isn't very evenly distributed yet.

Sharad Sharma is an entrepreneur-in-residence at Canaan Partners. He has a BE in electrical engineering from the Delhi College of Engineering. Sharma examines the transformation challenges facing the industry in his blog Orbit Change Conversations (http://orbitchange.com/blog/). Contact him at sharad_sharma@yahoo.com.

Future Imperfect

Vinton G. Cerf
Google

As the second decade of the 21st century dawns, predictions of global Internet digital transmissions reach as high as 667 exabytes (10^{18} bytes; http://en.wikipedia.org/wiki/SI_prefix#List_of_SI_prefixes) per year by 2013 (see http://telephonyonline.com/global/news/cisco-ip-traffic-0609/). Based on this prediction, traffic levels might easily exceed many zettabytes (10^{21} bytes, or 1,000 exabytes) by the end of the decade. Setting aside the challenge of somehow transporting all that traffic and wondering about the sources and sinks of it all, we might also focus on the nature of the information being transferred, how it's encoded, whether it's stored for future use, and whether it will always be possible to interpret as intended.

Storage Media

Without exaggerating, it seems fair to say that storage technology costs have dropped dramatically over time. A 10-Mbyte disk drive, the size of a shoe box, cost US$1,000 in 1979. In 2010, a 1.5-Tbyte disk drive costs about $120 retail. That translates into about 10^4 bytes/$ in 1979 and more than 10^{10} bytes/$ in 2010. If storage technology continues to increase in density and decrease in cost per Mbyte, we might anticipate consumer storage costs dropping by at least a factor of 100 in the next 10 years, suggesting petabyte (10^{15} bytes) disk drives costing between $100 and $1,000. Of course, the rate at which data can be transferred to and from such drives will be a major factor in their utility. Solid-state storage is faster but also more expensive, at least at present. A 1-Gbyte solid-state drive was available for $460 in late 2009. At that price point, a 1.5-Tbyte drive would cost about $4,600. These prices are focused on low-end consumer products. Larger-scale systems holding petabyte- to exabyte-range content are commensurately more expensive in absolute terms but possibly cheaper per Mbyte. As larger-scale systems are contemplated, operational costs, including housing, electricity, operators, and the like, contribute increasing percentages to the annual cost of maintaining large-scale storage systems.

The point of these observations is simply that it will be both possible and likely that the amount of digital content stored by 2010 will be extremely large, integrating over government, enterprise, and consumer storage systems. The question this article addresses is whether we'll be able to persistently and reliably retrieve and interpret the vast quantities of digital material stored away in various places.

Storage media have finite lifetimes. How

many 7-track tapes can still be read, even if you can find a 7-track tape drive to read them? What about punched paper tape? CD-ROM, DVD, and other polycarbonate media have uncertain lifetimes, and even when we can rely on them to be readable for many years, the equipment that can read these media might not have a comparable lifetime. Digital storage media such as thumb drives or memory sticks have migrated from Personal Computer Memory Card International Association (PCM-CIA) formats to USB and USB 2.0 connectors, and older devices might not interconnect to newer computers, desktops, and laptops. Where can you find a computer today that can read 8" Wang word processing disks, or 5 1/4" or 3 1/2" floppies? Most likely in a museum or perhaps in a specialty digital archive.

Digital Formats

The digital objects we store are remarkably diverse and range from simple text to complex spreadsheets, encoded digital images and video, and a wide range of text formats suitable for editing, printing, or display among many other application-specific formats. Anyone who has used local or remote computing services, and who has stored information away for a period of years, has encountered problems with properly interpreting the stored information. Trivial examples are occurring as new formats of digital images are invented and older formats are abandoned. Unless you have access to comprehensive conversion tools or the applications you're using continue to be supported by new operating system versions, it's entirely possible to lose the ability to interpret older file formats. Not all applications maintain backward compatibility with their own versions, to say nothing of ability to convert into and from a wide range of formats other than their own. Conversion often isn't capable of 100 percent fidelity, as anyone who has moved from one email application to another has discovered, for example. The same can be said for various word processing formats, spreadsheets, and other common applications.

How can we increase the likelihood that data generated in 2010 or earlier will still be accessible in useful form in 2020 and later? To demonstrate that this isn't a trivial exercise, consider that the providers of applications (whether open source or proprietary) are free to evolve, adapt, and abandon support for earlier versions. The same can be said for operating system providers. Applications are often bound to specific operating system versions and must be "upgraded" to deal with changes in the operating environment. In extreme cases, we might have to convert file formats as a consequence of application or operating system changes.

If we don't find suitable solutions to this problem, we face a future in which our digital information, even if preserved at the bit and byte level, will "rot" and become uninterpretable.

Solution Spaces

Among the more vexing problems is the evolution of application and operating system software or migration from one operating system to another. In some cases, older versions of applications don't work with new operating system releases or aren't available on the operating system platform of choice. Application providers might choose not to support further evolution of the software, including upgrades to operate on newer versions of the underlying operating system. Or, the application provider might choose to cease supporting certain application features and formats.

If we don't find suitable solutions to this problem, we face a future in which our digital information will "rot" and become uninterpretable.

If users of digital objects can maintain the older applications or operating environments, they might be able to continue to use them, but sometimes this isn't a choice that a user can make. I maintained two operational Apple IIe systems with their 5 1/4" floppy drives for more than 10 years but ultimately acquired a Macintosh that had a special Apple IIe emulator and I/O systems that could support the older disk drives. Eventually, I copied everything onto newer disk drives and relied on conversion software to map the older file formats. This worked for some but not all of the digital objects I'd created in the preceding decade. Word processing documents were transferable, but the formatting conventions weren't

directly transformable between the older and newer word processing applications. Although special-purpose converters might have been available or could have been written — and in some cases were written — this isn't something we can always rely on.

If the rights holder to the application or operating system in question were to permit third parties to offer remote access in a cloud-based computing environment, it might be possible to run applications or operating systems that developers no longer supported. This kind of licensing would plainly require creative licensing and access controls, especially for proprietary software. If a software supplier goes out of business, we might wonder about provisions for access to source code to allow for support in the future, if anyone is willing to provide it, or acquisition by those depending on the software for interpretation of files of data created with it. Open source software might be somewhat easier to manage from the intellectual property perspective.

Digital Vellum

Among the most reliable and survivable formats for text and imagery preservation is vellum (calf, goat, or sheep skin). Manuscripts prepared more than a thousand years ago on this writing material can be read today and are often as beautiful and colorful as they were when first written. We have only to look at some of the illuminated manuscripts or codices dating from the 10th century to appreciate this. What steps might we take to create a kind of digital vellum that could last as long as this or longer?

Adobe Systems has made one interesting attempt with its PDF archive format (PDF/A-1; www.digitalpreservation.gov/formats/fdd/fdd000125.shtml) that the ISO has standardized as ISO 19005-1. Widespread use of this format and continued support for it throughout Adobe's releases of new PDF versions have created at least one instance of an intended long-term digital archival format. In this case, a company has made a commitment to the notion of long-term archiving. It remains an open question, of course, as to the longevity of the company itself and access to its software. All the issues raised in the preceding section are relevant to this example.

Various other attempts at open document formats exist, such as OpenDocument format 1.2 (and further versions) developed by OASIS (see www.oasis-open.org). The Joint Photographic Experts Group has developed standards for still imagery (JPEG; www.jpeg.org), and the Motion Pictures Experts Group has developed them for motion pictures and video (MPEG; www.mpeg.org). Indeed, standards in general play a major role in helping reduce the number of distinct formats that might require support, but even these standards evolve with time, and transformations from older to newer ones might not always be feasible or easily implemented. The World Wide Web application on the Internet uses HTML to describe Web page layouts. The W3C is just reaching closure on its HTML5 specification (http://dev.w3.org/html5/spec/Overview.html). Browsers have had to adapt to interpreting older and newer formats. XML (www.w3.org/XML/) is a data description language. High-level language text (such as Java or JavaScript; see www.java.com/en/ and www.javascript.com) embedded in Web pages adds to the mix of conventions that need to be supported. Anyone exploring this space will find hundreds if not thousands of formats in use.

Finding Objects on the Internet

Related to the format of digital objects is also the ability to identify and find them. It's common on the Internet today to reference Web pages using Uniform Resource Identifiers (URIs), which come in two flavors: Uniform Resource Locators (URLs) and Uniform Resource Names (URNs). The URL is the most common, and many examples of these appear in this article. Embedded in most URLs is a domain name (such as www.google.com). Domain names aren't necessarily stable because they exist only as long as the domain name holder (also called the *registrant*) continues to pay the annual fee to keep the name registered and resolvable (that is, translatable from the name to an Internet address). If the registrant loses the registration or the domain name registry fails, the associated URLs might no longer resolve, losing access to the associated Web page. URNs are generally not dependent on specific domain names but still need to be translated into Internet addresses before we can access the objects.

An interesting foray into this problem area is called the Digital Object Identifier (DOI; www.doi.org), which is based on earlier work at the Corporation for National Research Initiatives

(www.cnri.reston.va.us) on digital libraries and the Handle System (www.cnri.reston.va.us/doa.html) in particular. Objects are given unique digital identifiers that we can look up in a directory intended to be accessible far into the future. The directory entries point to object repositories where the digital objects are stored and can be retrieved via the Internet. The system can use but doesn't depend on the Internet's Domain Name System and includes metadata describing the object, its ownership, formats, access modes, and a wide range of other salient facts.

As we look toward a future filled with an increasingly large store of digital objects, it's vital that we solve the problems of long-term storage, retrieval, and interpretation of our digital treasures. Absent such attention, we'll preside over an increasingly large store of rotting bits whose meaning has leached away with time. We can hope that the motivation to circumvent such a future will spur creative solutions and the means to implement them.

Vinton G. Cerf is vice president and chief Internet evangelist at Google. His research interests include computer networking, space communications, inter-cloud communications, and security. Cerf has a PhD in computer science from the University of California, Los Angeles. Contact him at vint@google.com.

Warehouse-Scale Computers

Urs Hölzle and Luiz André Barroso
Google

Computing is shifting away from desktops and small isolated servers and toward massive data centers accessed by small client devices. These large data centers are an emerging class of machines themselves, which we call warehouse-scale computers (WSCs).[1]

For example, consider an Internet service in which each user request requires thousands of binary instances from tens of individual programs to work in a coordinated fashion, with the hardware being a collection of WSCs distributed around the globe. The design, programming, and operation of this new machine class will be among the most challenging technical problems computer scientists face in the coming decade. At the surface, it looks easy — just a bunch of servers in a building, right?

Wrong. Here's a sampling of the problems we must solve to make WSCs ubiquitous.

Reliability

At scale, everything will fail. In a cluster of 10,000, even servers that fail only once every 30 years will fail once a day. If you store petabytes of data, you'll find bit errors that the hardware's error-detection mechanisms won't catch.

Thus, any application running on thousands of servers must deal automatically with failures (including ones you don't usually think of) and consider fault recovery a permanent background activity.

Availability

Achieving high performance and high availability in such a failure-vulnerable system requires consistency compromises.[2–3] In other words, you can have a system with either strong atomicity, consistency, isolation, and durability (ACID) guarantees or high availability, but not both.

In large-scale storage systems, availability is paramount, so consistency guarantees are weaker than in databases. This weakness can result in more complex services or APIs. For Internet services — and large-scale computing as a whole — to thrive, it's not enough to simply overcome these complexity challenges. We must solve them in a way that's consistent with high product innovation and programmer productivity. In other words, we must hide the complexity low enough in the architecture that application developers are not burdened by it and are free to quickly develop and test new product ideas.

MapReduce makes the comparatively simple case of parallel batch applications easy to program,[4] but no similarly simple solution exists yet for online applications with database-like needs.

Cost

Many warehouse-scale applications implement advertising-supported consumer Internet services, a business model with low per-user annual revenues. WSCs must therefore run cheaply.

They can't depend on aggressive replication or high-end hardware to achieve reliability.

Similarly, the administrative costs of running such services must be much lower than what is typical for IT services.

Latency and Locality

Any given data structure could reside in a local on-chip cache, a disk drive across the ocean, or somewhere in between. A large, cost-efficient Internet service must orchestrate data movement and distribution across an increasingly wide range of storage technologies and locality domains, taking into account user location, network costs, failure domains, and application needs. For simplicity, it would be best to manage data location automatically, but large discrepancies in data-access speeds and application-failure demands make this hard to do in practice.

Although the rate of innovation in Web-based services and applications is remarkable, we've only taken the first steps in exploiting this new model.

Data Center Efficiency

You've probably heard that data centers consume lots of energy and typically waste 1 to 2 watts of power for every watt the computing equipment consumes. On average, unfortunately, that's true. But improved building-level efficiency is no longer a research problem because efficient data centers do exist.

Google's data centers, for example, consume less than 20 percent energy than the servers use.[5] Bringing the average data center to this efficiency level and further reducing that energy overhead remain important challenges.

Server Efficiency

Servers themselves can be energy hogs, losing substantial energy in power conversions and when running at low to medium utilization levels. Efficient power conversions are practical today (for example, see the Climate Savers Computing Initiative at www.climatesaverscomputing.org), but they aren't consistently implemented yet.

The hard problem is to make data centers *energy-proportional*,[6] which means using 10 percent of the maximum power when the system is only 10 percent busy. (Today's servers consume roughly 50 to 60 percent of their maximum power even when idle.) Energy proportionality is important because most server farms spend much of their time well below maximum utilization levels.

Increasing Parallelism

Single-core speeds have improved only slowly over the past decade. Speed improvements have come mostly from increasing the number of cores per chip. The semiconductor industry expects these trends to continue. However, WSC workloads are growing at least as fast as Moore's law, so we must parallelize the workloads to scale their processing.

Even though large warehouse-scale applications tend to be easier to parallelize than some others, Amdahl's Law still rules: When the sequential processing speed ceases to improve, the burden to find more concurrency increases, which in turn increases the difficulty of implementing scalable infrastructures.

Although the rate of innovation in Web-based services and applications is already remarkable, we've only taken the first steps in exploiting this new model. To realize its full potential, we must tackle unprecedented levels of scaling and programming complexity in systems design.

References

1. L.A. Barroso and Urs Hölzle, "The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines," *Synthesis Lectures on Computer Architecture #6*, Morgan & Claypool Publishers, 2009, pp. 1–108; www.morganclaypool.com/doi/abs/10.2200/S00193ED1V01Y200905CAC006.
2. A. Fox et al., "Cluster-Based Scalable Network Services," *Proc. 16th ACM Symp. Operating Systems Principles* (SOSP 97), ACM Press, 1997, pp. 78–91.
3. G. DeCandia et al., "Dynamo: Amazon's Highly Available Key-Value Store," *Proc. 21st ACM Symp. Operating Systems Principles* (SOSP 07), ACM Press, 2007, pp. 205–220.
4. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Comm. ACM*, Jan.

2008, pp. 107–113.
5. "Data Center Efficiency Measurements," white paper, Google, 2009; www.google.com/corporate/green/datacenters/measuring.html.
6. L.A. Barroso and Urs Hölzle, "The Case for Energy-Proportional Computing," *Computer*, Dec. 2007, pp. 33–37.

Urs Hölzle is the senior vice president of operations at Google and a Google Fellow. His interests include large-scale clusters, cluster networking, Internet performance, and data center design. Hölzle has a PhD in computer science from Stanford University. Contact him at urs@google.com.

Luiz André Barroso is a distinguished engineer at Google. His interests include distributed systems, energy efficiency, and the design of warehouse-scale computers. Barroso has a PhD in computer engineering from the University of Southern California. Contact him at luiz@google.com.

The Internet of Things: Here Now and Coming Soon

Geoff Mulligan
IPSO Alliance

A transformation is coming to the Internet that will enhance our personal lives and forge advances in energy conservation, healthcare, home safety, environmental monitoring, and countless other facets of our world. The Internet of Things was first mentioned in work done at MIT in 1999 related to research into RFID tags. The concept was and is about connecting the physical world — things — to networks and tying them all together with the Internet. Vint Cerf, Google's chief Internet evangelist, describes it this way: "The Internet of the future will be suffused with software, information, data archives, and populated with devices, appliances, and people who are interacting with and through this rich fabric" (http://googleblog.blogspot.com/2008/09/next-internet.html). Although it was just an idea when we entered this century, it's now becoming a reality, and over the next decade, we'll see things we never thought could be on the Internet getting connected, with profound impacts to computing, protocols, socialization, privacy, and our lives.

Smart Objects

Smart objects — any device that combines local processing power with communications capabilities — are a reality. For years, the idea of pushing IP (the Internet Protocol) into small 8- and 16-bit devices with tens of Kbytes of program storage and possibly operating on battery power was thought to be impossible: the code would be too big, the packets too large, and the protocol too heavy for these low-speed, low-power networks. IP implementations now exist that run on sub-$2 micro-controllers with as little as 16 Kbytes of flash memory and 4 Kbytes of RAM. These IP stacks statelessly compress a 40-byte IPv6 header down to just 3 bytes, allowing the efficient transfer of packets even with battery-operated RF, while still providing IP-level end-to-end integrity and security, both of which are lost when using proprietary protocol-translation gateways. Within the next two years, as the power consumption of micro-controllers and embedded radios continues to decrease and the efficiency of batteries, photo-voltaics, and energy harvesting increases, we'll see a crossover in which these wireless sensor devices will be "always on."

Today

IP-based wireless sensor and control networks are deployed throughout the industry today. More than one million IP-enabled electric meters deployed in 2009 now support automated meter reading, and hundreds of thousands of streetlights are interconnected with IP-based RF mesh networks to provide remote condition monitoring. IP-enabled temperature, humidity, and motion sensors are now installed in office buildings and connected with existing IP infrastructure to augment building control systems. IP- and RF-interconnected clocks are now used in hospitals to ensure precisely accurate times for medical events. In August 2009, a 61-year-old woman had surgery to install an IP-connected pacemaker; her doctors can now remotely check on her condition. Smart grid projects, energy management systems, and telemedicine portend even more pervasive use of smart objects, especially IP smart objects, which are the building blocks for tomorrow's Internet of Things.

Tomorrow

"When the parents are away, the children will play," or so goes an old saying, but wouldn't

it be nice to know they're safe while you're away? New applications will allow appliances such as the stove to alert parents that they've been turned on (or were left on) and let you turn them off remotely. When you forget to turn off the lights, you can rectify that situation remotely. Smoke detectors will "talk" to gas appliances to shut them off when an alarm sounds and then send an alert to your phone. For people with parents far from home, appliances could watch for typical usage patterns (refrigerator door opening and closing, oven or microwave being used, and motion throughout the house) and message you if these events aren't occurring as expected.

In addition to safety and security, the Internet of Things will mean enhanced convenience. Rather than needing to figure out how to heat a meal in the microwave, the microwave will read the RFID tag on the container and request heating instructions from the manufacturer via the Internet — all you'll have to do is press "cook." When you check in at your hotel, your unique preferences will be sent to the room so that the temperature is correct, lights lit, and radio preprogrammed. Within your home, rather than having just a single temperature sensor (your thermostat), temperature sensors can be set throughout the house along with occupancy sensors to ensure that the rooms you actually use most stay at the requested temperature. Parking spaces can send messages indicating availability either around the next corner or on the next level in a parking structure, all relayed to your phone via a streetlight network. For eldercare and remote healthcare, you'll find that you can take blood pressure or glucose levels with an IP smart object and have the results sent to your doctor securely and automatically via an already installed home network.

Privacy

With all this oversight and viewing into our daily affairs, we must be cognizant of the possibility of technical and ethical abuse and the need to protect our privacy. Not only do we need to ensure that security mechanisms and protocols are properly designed but also properly used and defined for data usage and ownership. Who owns the information about home usage of appliances or products — us, the utility, the appliance manufacturer, the warranty service company, or maybe all of these, in various contexts? With motion sensors in our homes, cars, and phones able to report our location, can a thief check to see if we're home? Ethically, if our microwave reports the foods we eat, should our doctor or insurance company know that we just ate an entire bag of "theater butter" microwave popcorn?

With these billions, or billions of billions, of devices coming online, we must find ways to allow them to either be self-configuring or so easily configured that anyone can do it. Mark Weiser, widely considered to be the father of ubiquitous computing, said, "the most profound technologies are those that disappear ... they weave themselves into the fabric of everyday life until they are indistinguishable from it."[1] Although the protocols of today such as stateless address auto configuration and 6lowpan help get us closer, they don't completely solve problems nor enable completely self-forming, self-healing ad hoc networks. Additionally, these new Internet objects must be able to "learn" what servers and services they can and, more importantly, should talk to. They must be able to advertise the services and data that they can provide so that they can seamlessly participate in the Semantic Web. New transport and application protocols and data formats must also be defined for these embedded and nearly invisible devices.

The enhanced connectivity between devices in the Internet of Things is expressly designed to engage us in making informed decisions about creating a safer, greener, healthier, more efficient, and far less wasteful world. The Internet of Things will provide nearly limitless amounts of information and a much higher granularity of measurement, but we need to be ready for this explosion of data and control. Yesterday's Internet = "anytime, anyplace, anyone." Today's Internet = "anytime, anyplace, anything."

Reference

1. "The Computer for the 21st Century," *Scientific Am.*, vol. 265, no. 9, 1991, pp. 66–75.

Geoff Mulligan is chairman of the IPSO Alliance, president of Proto6 (an Internet technology consulting firm), and chair of the IETF 6lowpan working group. His research interests include embedded IP sensor networking, computer and network security, and IPv6. Mulligan has an MS in computer science from the University of Denver. Contact him at geoff@proto6.com.

Interplanetary Internetworking

Adrian Hooke
NASA Headquarters

The 53 years since Sputnik-1 opened the Space Age have brought extraordinary advances in deep-space communications. Data rates, in particular, have increased from a few bits per second in low Earth orbit to multimegabits per second from Mars. The earliest communication systems transmitted spacecraft telemetry and telecommand information as analog signals, derived from mechanical rotating commutators that sampled key measurements and modulated them directly onto the radio carrier.

As digital communications emerged in the early 1960s, electronic equivalents of the old commutators were implemented. These systems sampled each spacecraft measurement in a fixed sequence, which forced onboard payloads to communicate over a synchronous time slot in the transmitted stream, often resulting in over- or undersampling.

In the late 1970s, this communications model began shifting to new packetized data-transfer technology that let each onboard application create and consume information asynchronously. An autonomous "space packet" encoded a complete measurement set for transmission at a data rate appropriate to a specific investigation. The onboard spacecraft data system then switched the packets associated with different applications in and out of the radio channel connecting the spacecraft with the Earth. An application-process ID (APID) tagged each packet as belonging to a single information flow between an onboard application and one or more ground users.

An International Enterprise

By 1982, the number of countries embarking on space missions had grown, and the international space community began considering ways to share their mission-support infrastructure to facilitate collaborative space exploration. The Consultative Committee for Space Data Systems (CCSDS) was therefore formed to address the need for new data-handling approaches that would allow spacecraft and ground infrastructure from different organizations to interoperate.

Building on the new packetized data-transmission technology, the CCSDS produced the first generation of international standards for packet telemetry and telecommand. Simple data routing was implemented based on the space packet's APID. Although the APID lacks a fully formed source- and destination-addressing system and is therefore hardly the foundation of a fully fledged internetworking protocol, it has nevertheless served the community well for almost three decades as the workhorse for relatively simple data transfer between a single spacecraft and its ground support system. To date, almost 450 space missions have adopted the CCSDS packetized architecture.

By the mid 1980s, the CCSDS was vigorously pursuing standardization of an expanded set of space-to-ground data-communications techniques, including advanced modulation, channel coding, and data compression. When plans emerged to build the International Space Station (ISS), special CCSDS working groups began addressing its unique challenges of high forward-data rates, very high return-data rates, and many different traffic types — including audio and video to support flight crew activities.

The participation of multiple space agencies from Europe, Japan, and the US in the ISS program resulted in a more complex space configuration, involving several cooperating space vehicles and, consequently, the need for more powerful internetworking techniques to transfer user information between the spacecraft and the ground. In the mid 1980s, open systems interconnection (OSI) was in full swing, and the International Standards Organization (ISO) was developing a suite of new standards to follow the OSI seven-layer network model. The CCSDS therefore updated its packetized data-transmission protocols to support internetworking traffic across the space-to-ground radio links using the Connectionless Network Layer Protocol (ISO 8473). A decade later, the ISS had gone through extensive redesigns, but the updated CCSDS standard remained the bedrock of international interoperability.

Meanwhile, in the terrestrial data-communications community, the ISO protocol suite gave way to the emerging Internet protocol suite based on IP rather than ISO 8473. Today, the ISS still runs the CCSDS protocols, but its space-to-ground traffic is increasingly IP-based.

A Delay-Tolerant Internet

The experience with early ISS internetworking approaches led the CCSDS to experiment

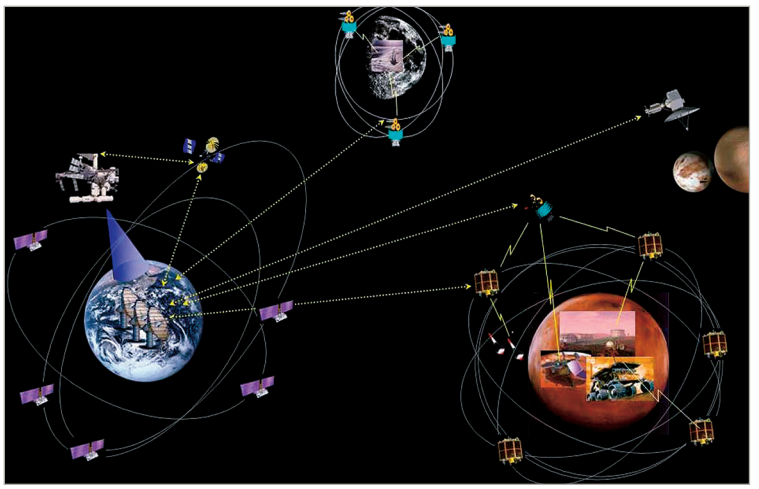


Figure 1. Interplanetary internetworking. The delay-tolerant network (DTN) protocol suite that will support communications among the increasing number of vehicles in free space and on other solar system bodies has already been demonstrated in deep space. (Source: NASA, 2009; used with permission.)

with adapting and extending the terrestrial TCP/IP suite to better match the long delays and intermittent connectivity characteristic of space communications. In particular, the CCSDS designed some TCP extensions that have become the foundation of many performance-enhancing proxies now widely deployed across commercial and military satellite-communications communities.

Building on that experience, the CCSDS and terrestrial Internet communities began working more closely together. In 1999, DARPA (which funded much of the early Internet development) independently allocated resources from its Next-Generation Internet program to study the possible architecture of an Interplanetary Internet. In parallel with the DARPA study, CCSDS developed a CCSDS File Delivery Protocol (CFDP) to allow bidirectional space-ground file transfer over long-delay links with asymmetric data-transmission capabilities. Because remote spacecraft are often out of Earth's view and rarely have a contemporaneous end-to-end data path to the ground, the CFDP implemented new store-and-forward techniques to support communications when the path is interrupted.

Delay and disruption tolerance (neither of them hallmarks of the IP suite) therefore emerged as the key characteristics needed for interplanetary internetworking. The CCSDS and DARPA work consequently started converging on a new delay-tolerant networking (DTN) approach to communicating in difficult environments. Recognizing the potential for these new techniques to also extend the terrestrial Internet's reach into areas with underprovisioned or highly stressed communications resources, the Internet Research Task Force formed a DTN research group to advance the new technology.

The DTN architecture that emerged from this group is an overlay network based on a new Bundle Protocol (BP). The relationship of BP to IP is as IP to Ethernet — that is, an IP-based network connection (perhaps the entire Internet) can be one "link" in a DTN end-to-end data path. The Interplanetary Internet is effectively a "network of Internets." The BP itself sits below the end-to-end space applications and above the individual subnetworks, which might be the terrestrial Internet, individual long-haul CCSDS backbone space links, or local area networks at remote locations in space.

As we enter a new decade, space exploration is ready to exploit powerful internetworking capabilities. Mars already has a rudimentary network composed of proximity-communications payloads aboard orbiting spacecraft, which act as relays between roving vehicles and the Earth. Since becoming the primary communications path for Martian surface operations, these relays have transmitted roughly 95 percent of all rover data at much higher rates and with lower power requirements than the previous direct-to-Earth approach. Despite relatively short flyover contact durations, the high-rate communications have significantly increased the volume of end-to-end data delivery.

Looking ahead to missions launching to the Moon and Mars from 2015 onwards, we can expect international cooperation in executing complex missions to continue (see Figure 1). As the number of space vehicles grows — in free space and on other solar system bodies, so will the need to share data-transmission capabilities and to standardize intervehicular operations. The DTN protocol suite that can support this exciting new era is almost ready for deployment. NASA has already demonstrated DTN-based internetworking operating in deep space on the Epoxi spacecraft, and a permanent DTN node is now orbiting the Earth on the ISS. The CCSDS standardization of DTN has begun, and

a common set of flight-qualified DTN software is in development.

For this magazine's 2020 forecast issue, I hope to report the planned expansion of Interplanetary Internet operations. Then, who knows how fast it will grow?

Adrian Hooke is manager of space networking architecture, technology, and standards in the NASA Headquarters' Space Communications and Navigation office. His research interests include extending the terrestrial Internet Protocol suite into space. Hooke has a BS with honors in electronic and electrical engineering from the University of Birmingham, UK. He chairs the CCSDS Engineering Steering Group and the US Technical Advisory Group to ISO TC20/SC13, Space Data and Information Transfer Systems. Contact him at adrian.j.hooke@jpl.nasa.gov.

GENI: Opening Up New Classes of Experiments in Global Networking

Chip Elliott
Global Environment for Network Innovations

The Global Environment for Network Innovations (GENI) is a suite of research infrastructure components rapidly taking shape in prototype form across the US. It is sponsored by the US National Science Foundation, with the goal of becoming the world's first laboratory environment for exploring future Internets at scale, promoting innovations in network science, security, technologies, services, and applications.

GENI will allow academic and industrial researchers to perform a new class of experiments that tackle critically important issues in global communications networks:

- *Science issues*. We cannot currently understand or predict the behavior of complex, large-scale networks.
- *Innovation issues*. We currently face substantial barriers to innovation with novel architectures, services, and technologies.
- *Society issues*. We increasingly rely on the Internet but are unsure that we can trust its security, privacy, or resilience.

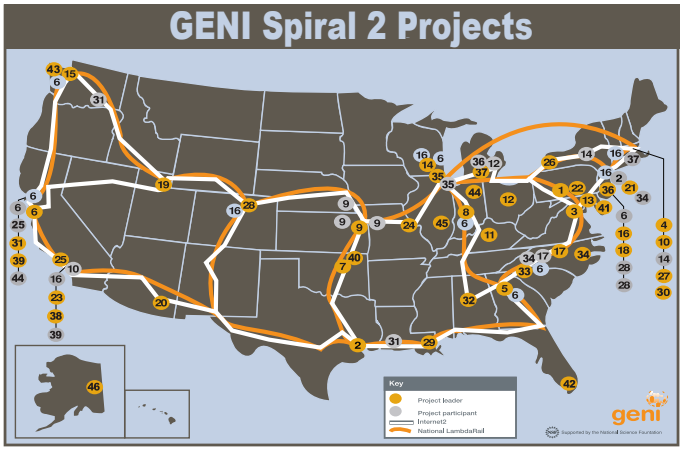
It will support two major types of experiments. The first is controlled and repeatable experiments, which will greatly help improve our scientific understanding of complex, large-scale networks. The second type is "in the wild" trials of experimental services that ride atop or connect to today's Internet and that engage large numbers of human participants. GENI will provide excellent instrumentation for both forms of experiments, as well as the requisite data archival and analysis tools.

Building GENI via Rapid Prototyping

GENI is being created as a series of rapid prototypes via spiral development so that hands-on experience with early experimentation and trials can drive its evolution. Many leading researchers are engaged in planning and prototyping GENI, including those who have created PlanetLab, Emulab, OpenFlow, the Orbit and Cyber-Defense Technology Experimental Research (Deter) testbeds, and a variety of other innovative research tools. Industrial research teams are also engaged with these academic teams, including AT&T, CA Labs, HP Labs, IBM, NEC, and Sparta, as are Internet2 and NLR, the two US national research backbones, and several regional optical networks.

Rather than build a separate, parallel set of infrastructure "as big as the Internet," which is clearly infeasible, current plans call for GENI-enabling existing testbeds, campuses, regional and backbone networks, cloud computation services, and commercial equipment. GENI can then incorporate these networks and services by federation, rather than constructing and operating a separate infrastructure for experimental research.

"At-scale" experimentation, as currently envisioned, may ultimately grow to involve tens or hundreds of thousands of human participants and computers, and thus needs a way by which such experiments may be smoothly migrated out of the GENI infrastructure and into production use as new services. Starting in October 2009, the GENI project will begin to pave the way to such experiments by a meso-scale build-out through more than a dozen US campuses, two national backbones, and several regional networks. If this effort proves successful, it will provide a path toward more substantial build-out.



Project Name	Project Lead	Project Participants
1. CMUlab	Carnegie Mellon University	
2. D Meas, LEARN	University of Houston	Columbia University
3. Digital Object Registry	Corporation for National Research Initiatives (CNRI)	
4. CLOUD-CTL, DOME, ViSE	University of Massachusetts Amherst	
5. DTunnels	The Georgia Institute of Technology	
6. EnterpriseGENI, OpenFlow	Stanford University	Princeton University University of California, Berkeley
	Clemson University	
	Georgia Institute of Technology	
	Indiana University	
	Nicira Networks	
	Princeton University	
	Rutgers University	
	University of Wisconsin	
	University of Washington	
7. GENI4YR	Langston University	
8. GMOC, netKarma, K-GENI	Indiana University	
9. GpENI	University of Kansas	Kansas State University, University of Nebraska-Lincoln
	The University of Missouri-Kansas City	
10. GushProto	Williams College	UC San Diego
11. INSTOOLS, ISM Infrastructure	University of Kentucky	
12. KANSEI, OTM	Ohio State University	Wayne State University
13. MAX	University of Maryland	
14. MeasurementSys	University of Wisconsin-Madison	Boston University Colgate University
15. MillionNodeGENI, Security	University of Washington	
16. ORBIT, WiMAX	Rutgers University	UCLA, Los Angeles, CA
	Columbia University, NY, NY	University of Colorado, Boulder, CO
	Polytechnic University of NYU, Brooklyn, NY	University of Massachusetts, Amherst University of Wisconsin, Madison, WI
17. ORCA/BEN	The Renaissance Computing Institute (RENCI)	Duke University
18. PlanetLab, Scaffold, Federation	Princeton University	Universite Pierre et Marie Curie (UPMC)
19. ProtoGENI	University of Utah	
20. PROVSERV	University of Arizona	
21. ERM	Columbia	
22. REGOPT	Pittsburgh Supercomputing Center (PSC)	
23. SECARCH, Distributed Identity	SPARTA, Inc.	
24. SPP	Washington University	
25. TIED	USC Information Sciences Institute	University of California, Berkeley
26. UB_OANets	SUNY Buffalo	
27. UMLPEN	University of Massachusetts Lowell	
28. CR-GENI	University of Colorado Boulder	Radio Technology Systems LLC Rutgers University
29. CRON-T	Louisiana State University	
30. Design of Information Subs	MIT	
31. DSL, HIVE	UC Davis	Batelle CA Labs
32. EXP-SEC	University of Alabama	
33. FPGA-RADIO	Clemson University	
34. GENI IMF	North Carolina State University	The Renaissance Computing Institute (RENCI) Columbia University
35. iGENI	Northwestern University	University of Illinois Chicago
36. LAMP	University of Delaware	Internet2
37. LEFA, Supercharged Planetlab	Internet2	Brown University
38. NLR	Cypress, CA	
39. Open--CIRRUS	HP Labs, Palo Alto	UCSD
40. OKGems	Oklahoma State University	
41. PIGEON-NET	Howard University	
42. PrimoGENI	Florida International University	
43. QUILT	The Quilt	
44. S3-GENI	Purdue University	HP Labs
45. SEC-POL	University of Illinois (NCSA)	
46. VMI	University of Alaska Fairbanks	

Figure 1. Global Environment for Network Innovations (GENI) prototyping as of October 2009. As the figure shows, the mesoscale prototype is rapidly taking shape across the US.

Figure 1 shows current GENI participants, together with the two GENI-enabled national backbones that link their campuses.

Running New Classes of Experiments

Early experimentation will begin in 2010 and will drive GENI's evolving design. Within the

next 6 to 12 months, GENI will start to open up new areas of experimental research at the frontiers of network science and engineering — fields with significant potential for socioeconomic impact. These experiments may be fully compatible with today's Internet, variations or improvements on today's Internet protocols, or indeed radically novel "clean slate" designs.

Although research interests, and thus experiments, will evolve significantly over the coming decade, we expect the earliest types of GENI-based experiments to include the following areas.

Content Distribution Services

As the Internet is increasingly used to distribute high-bandwidth content (for example, video and virtual worlds), many researchers have focused on new, more scalable architectures for such services. GENI is well suited to such experiments, with its emphasis on deep programmability, clouds, and GENI-enabled campuses.

Disruption-Tolerant Networks

GENI is specifically aimed to enable large-scale, well-instrumented, repeatable experiments on novel protocols and architectures. Disruption-tolerant networks (DTNs) are a perfect case in point since many DTN architectures are independent of today's TCP/IP architecture. We expect several DTN experiments to begin on GENI within the coming months.

Measurement Campaigns

Researchers still lack basic knowledge and understanding of the Internet's behavior. GENI's emphasis on highly instrumented infrastructure will provide tools for capturing, analyzing, and sharing measurements on the global Internet as it evolves.

Novel Mobility Architectures

Many networking researchers have proposed novel protocols to improve support for mobile devices in the Internet architecture. GENI's near-term emphasis on wireless support throughout campuses allows real-world experimentation with these new protocols.

Novel Routing Architectures

As concerns have grown over the scalability of the global Internet routing architecture, particularly with the rise of multihoming, a number of research teams have proposed alternative global routing architectures. Although GENI will not be as big as the Internet, it may offer sufficient scalability so that such approaches can be tried out in a realistic, well-instrumented suite of infrastructure components.

Reliable Global Networks

As I mentioned before, all of us increasingly rely on the Internet but are increasingly uncertain that we can trust its security, privacy, or resilience. There is now growing interest in experimental efforts that will help ensure an Internet that is solid and reliable in the critically important role it now plays for society.

Virtualization Architectures

GENI's own architecture is based on end-to-end virtualization, which is now becoming an area of keen interest and study to networking researchers. Indeed, GENI prototyping teams are actively experimenting with new network architectures based on virtualization; in addition, we expect future virtualization experiments to run within the GENI infrastructure as it comes online.

Looking Ahead

GENI is a visionary project in its early stages; both opportunities and challenges abound. Its infrastructure is growing rapidly and will soon begin hosting a range of experiments in network science and engineering, with two overarching goals.

The first goal is *understanding*. We want to transform science in networking and distributed systems by

- enabling frontier research into the world's future sociotechnical networks;
- creating a strong interrelationship between theory and experiments in complex, large-scale network systems;
- greatly increasing the field's emphasis on large-scale, realistic experiments with widespread archiving, sharing, and analysis of experimental data; and
- exciting a new generation of students by providing a sense of shared adventure and exploration in a rapidly developing field.

The second goal is *innovation*. We want this field of research to have a high degree

of engagement with, and a strong impact on, industry, the economy, and our society. GENI can help achieve this by

- making up-to-the-minute technology broadly available for experimentation;
- stimulating new network services and architectures that support the nation's critical needs for security, privacy, and robust availability;
- dismantling barriers to entry so that individuals and small teams can rapidly extend innovation deep into the network core;
- encouraging large numbers of early adopters in the American public for cutting-edge experimental services produced by academia and industry;
- providing a graceful transition path from innovative research experiments to useful commercial service offerings; and
- emphasizing sustained, long-term research collaboration between academia and industry.

GENI is being designed and prototyped in an open, transparent process in which all may participate. We welcome engagement and participation (see www.geni.net) and encourage researchers to contact us with proposals for novel experiments. We are now increasingly confident that realistic, at-scale experimentation can help drive the next 10 years of our global communications infrastructure.

Chip Elliott is the principal investigator and project director for the Global Environment for Network Innovations (GENI). He's also Chief Engineer at BBN Technologies and a fellow of the American Association for the Advancement of Science and the IEEE, with more than 65 issued patents. Elliott has a BA in mathematics from Dartmouth College. Contact him at celliott@bbn.com.