



Imagining the Future: Thoughts on Computing

Daniel A. Reed, Dennis B. Gannon, and James R. Larus, *Microsoft*

New and compelling ideas are transforming the future of computing, bringing about a plethora of changes that have significant implications for our profession and our society and raising some profound technical questions.

Many bedrock assumptions in computing are fast crumbling as research ideas become everyday reality.

Processing power is migrating to a range of specialized devices, from ubiquitous smartphones to sensors embedded in everyday objects. Long-envisioned elements of natural user interfaces such as multitouch, 3D, and speech processing are entering the mainstream. Data storage density continues to increase at exponential rates, making it possible for individuals to maintain digital records of their lives and activities both locally and in the cloud. The Internet of Things—the growing and largely invisible web of interconnected smart objects—promises to transform the way we interact with everyday things. Not surprisingly, digital technology and our shifting expectations are challenging the traditional notions of privacy and security.

This morphing of new and old technologies poses exciting research and development challenges, offering the opportunity to develop compelling ideas and creative implementations that can transform the future. For example, the explosive growth of smartphones and wireless devices is creating a spectrum shortage. Can cognitive radio technology reshape the technical and business dynamics of communications into more nimble and adaptive spectrum usage?

As Figure 1 shows, projections of current trends suggest we'll see more than 50 billion Internet-connected devices in just a few years. How can we best manage and secure them? Can we reconcile our historical notions of privacy and security, rooted in person and place, with the new world of cloud services and transnational data flows? Each of today's cloud datacenters contain more computing and storage capacity than the entire Internet did just a few years ago. How can we best design these systems for resilient and energy-efficient operation?

New computing technologies also bring opportunities to help solve important societal problems. The population is aging in much of the world. Can these new technologies improve the quality of life for seniors? How will the revolutions in biology leverage advances in information technology to deliver truly personalized medicine? Can educational systems respond to the need for training people in the jobs of the future? Can digitally mediated education accommodate the need for just-in-time training and refreshing skills? Can we use these new technologies to manage the growing demand for clean energy more effectively?

Predicting the future is always fraught with peril, as any retrospective examination of technology predictions will show. However, these technological and societal changes are so significant that we must consider their effects on the future of research, computing education, and broader societal responses.

MAJOR COMPUTING TECHNOLOGY TRENDS

In computing, we've become accustomed to dizzying change, with orders-of-magnitude shifts in the capacities, performance, and costs of devices within just a few years. These changes accrue from both technological advances and the judicious integration of appropriate component

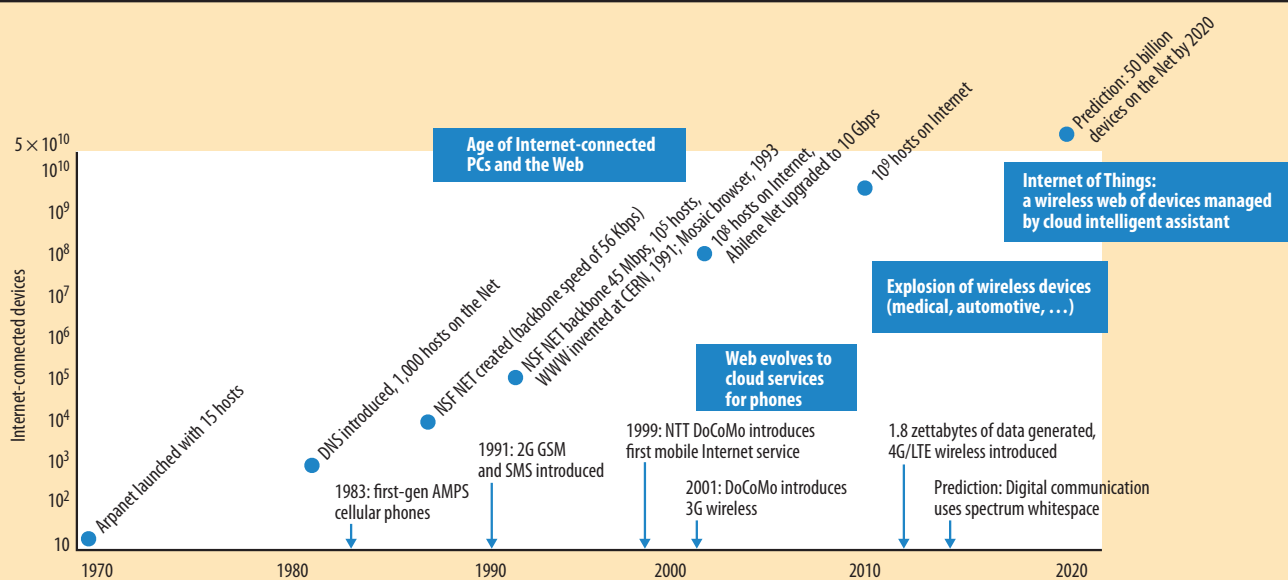


Figure 1. The Internet of Things will be possible because of advances in many technologies, including multifunction systems on chip, ubiquitous cloud services, and digital communication enabled by cognitive radio.

technologies. As the late Jim Gray wisely noted,¹ the ratios of component technologies' cost and performance determine the technical and economic feasibility of any design and its corresponding utility and application. In that spirit, it's instructive to consider some of the major component trends.

Silicon scaling and Moore's law

For decades, CMOS process scaling has provided faster processor chips, with lower power and cost via reductions in transistor feature size. This balanced, proportional change, called *Dennard scaling*,² is the technological enabler of Moore's law,³ and it is now ending due to simple physics: a transistor gate's insulator is now only a few atoms thick.

At this point, we can't continue to reduce the transistor supply voltage, which had declined by 30 percent with each reduction in feature size. Without significantly reducing the voltage, power density on chips grows quadratically as transistors shrink. Simply put, future process technology will allow the semiconductor industry to place more transistors on chips, but they won't be much faster, and without alternatives to standard CMOS, they won't be much more energy efficient either. An unfortunate implication is that it won't be possible to power all the transistors on a chip simultaneously.

This slowing of single-processor performance increases is already reshaping the computing landscape. Multicore processors have made parallel programming the norm rather than the exception. For parallel applications such as Web services and cloud computing, multicore chips bring immediate performance benefits. Moreover, multicore parallelism enables computational heterogeneity, with

specialized coprocessors performing specific tasks at far lower power, a boon to low-power and mobile devices.

Without doubt, advances in semiconductor technology will continue to drive improvements in multicore chip power, cost, and parallelism, enabling human interaction in richer and more natural ways, supported by large-scale cloud services and data analytics. A significant challenge will be creating the programming models and tools that facilitate software design in this new parallel and distributed world, and educating a new generation of students and developers in their use.

The many-device world, the Internet of Things, and the cloud

The history of computing is one of increasing democratization. Rare and expensive mainframe computers rapidly begat the first minicomputers, then workstations, and, more recently, personal computers. Today's smartphones and tablets are the latest generation of democratized access. Moore's law, increased integration, and high device volumes are bringing forth yet another transformation, the rising Internet of Things, with intelligence embedded in everyday objects.

Although computers have long been embedded in physical devices as controllers, the significant change is the ability to connect even the most inexpensive devices to the Internet. Each of us already owns dozens or hundreds of objects with embedded computing systems, and the number of "things" connected to the Internet greatly exceeds the number of people using them.

Cloud computing, which we define here as Internet-scale services hosted in massive datacenters, enables ubiquitous

Web searches, hosted software, and social networks. It also provides the analytics that enable mobile devices to adapt and personalize their behavior, for example, by using their location to find desirable nearby restaurants.

The cloud is the glue that binds the Internet of Things together. Consider an electric car that requires regular recharging. If every electric vehicle owner began recharging right after work, the power demand could potentially exceed the current electric grid's capacity. But the amount of charge a vehicle needs—and by when—often depends on where the driver must be the following day, information that could be inferred from a calendar, subject to privacy and security constraints. A driver's software agent could access the driver's calendar in the cloud, contact the vehicle to determine remaining battery capacity, and then negotiate with the power company's agent to schedule a charge. This cooperation, made possible by ubiquitous networks, shared data, and cloud-based agents, provides obvious benefits in regulating the load on the grid and increasing the appeal of electric vehicles by making them simpler to own.

The cloud also offers a platform for building sophisticated services that make Internet-connected devices far more than up-to-date replacements for the previous generation of “dumb” devices. The cloud can store data that needs to be always accessible to a large number of separate devices, and it provides computing resources sufficient for sophisticated agents. The computers embedded in devices will be limited by cost, power, and size constraints, which in turn will bound the versatility and sophistication of the software that can run directly on them. Cloud-based agents eliminate many of these restrictions.

Naturally, this shift raises many challenges in security, privacy, and autonomy, particularly as these network-connected devices are open to malicious attacks as well as more insidious invasions of privacy. Although improvements in hardware and software can raise better barriers to malicious attacks, privacy requires broader agreement on societal norms, as well as mechanisms for enforcing them.

Cloud datacenters and scaling

At the end of 2011, the amount of information created and replicated surpassed 1.8 zettabytes,⁴ and the production rate continues to increase. The implication of this growth is staggering: 90 percent of the data in the world today was created in the past two years alone. This data comes from myriad sources, including the explosive amount of scientific and engineering information gleaned from inexpensive, high-resolution sensors; digital communications and computer-mediated human interactions; intelligent infrastructure in buildings and transportation systems; and industrial machines that sense, create, and communicate data via the Internet of Things.

This torrent of digital data typically resides in a globally distributed network of cloud datacenters, with the largest consuming tens of megawatts of power to support hundreds of thousands of servers and many tens of petabytes of storage. Microsoft, Google, Amazon, Facebook, and other cloud service operators are building worldwide networks of these datacenters to host applications and deliver consumer services. Just as mobile devices and the Internet of Things have driven innovation in packaging and power management, so too has the construction of cloud datacenters.



The cloud can store data that needs to be always accessible to a large number of separate devices, and it provides computing resources sufficient for sophisticated agents.

The traditional datacenter, with its raised floor, fluorescent lighting, and circulating air, has given way to new models that emphasize energy efficiency, rapid deployment, and resilient operation. Datacenter building blocks are now based on shipping containers, each preconfigured with thousands of servers and needing only basic connections for power, networking, and cooling. They can be transported and deployed quickly, separating facility construction from equipment configuration. Similarly, traditional cooling models have given way to either higher-temperature operation, accepting the risk of slightly more component failures in exchange for reduced cooling costs, or airside economization based on using ambient air for cooling.

But despite improvements in their physical infrastructure, datacenters continue to raise difficult research issues in low-power node design, packaging, integration, and operation. Beyond hardware, the vast collection of machines and their huge workloads pose challenges in network management, fault tolerance and resilience, workload optimization and configuration, and system architecture.


Software and cloud services

Since the birth of modern digital computing, software has been an artifact that users developed or purchased along with a computer to accomplish a computational or data analysis task. Increasingly, though, consumers and businesses are buying the *service* provided by the software, rather than the software itself. It is now possible to write a paper or book using Microsoft Office 365 or Google Docs and store the document using Microsoft's Windows Azure or Amazon Web Services without buying and installing a single piece of software.

This trend holds great promise to simplify many human-computer interactions, replacing low-level abstrac-

tions such as machines and files with concepts such as a digital photograph album that is accessible anywhere and from many devices. Economic models have expanded as well, to include advertising for supported software service delivery, most notably Web search. This new paradigm, based on diverse devices bound together by shared storage and computation in the cloud, requires new software development practices and techniques.

Programming and evolving services atop a concurrent, distributed system that supports millions of geographically distributed users is a daunting challenge because the services must be highly available and reliable: their failure would disrupt millions of peoples' lives. Consequently, researchers must develop and deploy software updates and responses to errors or security attacks while the systems are in operation.



Cloud-based digital assistants could take on many of the same tasks as a human assistant while also providing access to the wealth of data stored in the cloud.

Out of necessity, cloud services platforms are already improving software development processes. Real-time feedback on errors⁵ now allows development teams to identify latent defects quickly, and the tight connection between a device and the cloud allows for rapid, continual improvements to deployed software. Nevertheless, software still continues to have large numbers of defects that often serve as portals for malicious attacks. Improved software architectures⁶ and wider adoption of software defect detection tools could do much to improve attack resilience, but more work is needed in this vital area.

Post-WIMP and NUI

Our traditional interaction with computing devices, whether PCs, tablets, or smartphones, is through GUIs that elicit well-defined system responses. We can trace the now-ubiquitous window, icon, menu, pointing (WIMP) paradigm that has dominated our interaction with computers for the past three decades to the pioneering work at Xerox PARC.⁷

The first glimmers of a new model of human-computer interaction based on natural user interfaces (NUIs) are now emerging. With a NUI, the user interacts with the computer through human gesture and speech. The idea has been around for a long time in science fiction literature—for example, the paranoid computer HAL in Stanley Kubrick's *2001: A Space Odyssey* had a well-developed NUI. The technology for speech recognition has a long history as well,⁸ but it has evolved rapidly over the past 10 years, with many smartphones now allowing speech input for Web search-

ers. A greater challenge has been the recognition of human gestures. Work on computer vision has been under way for many years,⁹ and gesture recognition has been intensively studied for the past 15 years.¹⁰

Microsoft's Kinect is the most recent commercial example of a hardware realization of NUI ideas. Kinect and Xbox both use a depth camera and a microphone array to read hand and body motions, but Kinect can also respond to spoken commands, such as, "Kinect, search videos for *Star Wars*," or, "Kinect, play the video." The breakthrough that made the Kinect possible¹¹ has been the use of depth images and a highly trained classifier that maps body parts to 3D models, thereby allowing human users to move, point, and gesture in natural ways that the system can interpret as computer commands.

Anticipatory, assistive computing

To achieve a truly natural user interface, we must move beyond reacting to explicit commands and build systems that understand implicit queries and anticipate questions and actions. Cloud-based digital assistants could take on many of the same tasks as a human assistant while also providing access to the wealth of data stored in the cloud. Apple's Siri is an excellent first look at this capability (www.apple.com/iphone/features). Such an assistant would be contextually aware of where users are and what they are doing by monitoring the stream of data from personal digital devices. This contextual information could also help the digital assistant handle ill-defined queries and problems.

Building computer systems that can anticipate actions or interests is an extremely difficult problem, but researchers are making progress. Search engines, for example, are becoming much better at identifying user intent—by mining search query data, they can display the most likely completions of the first few words of a search string. Likewise, ShadowDraw,¹² a new freehand drawing program, searches collections of images similar to a user's sketch to find a template that can help guide the drawing. Given the user's strokes, the system dynamically retrieves matching images, aligns them to the evolving drawing, and weights them based on a matching score, all in real time.

The rise of "big data analytics" that emerged from the data deluge now enables us to infer other user behavior patterns and (often) understand user intent. As seen on *Jeopardy!*, IBM's Watson is a breakthrough application of data mining in which the system applies dozens of specialized, individualized machine-learning algorithms in parallel to petabytes of unstructured textual data distributed across multiple processors. Watson can discover the implicit connection between seemingly unrelated words and phrases.

Multimodal communication

The rapid growth of Wi-Fi devices and smartphones has created unprecedented expectations for wireless data

access, with associated pressures on telecommunications regulators and cellular providers to meet demand via spectrum allocation and infrastructure, respectively. As anyone who has ever experienced a dropped call or been unable to connect to a Wi-Fi hotspot in a crowded area knows, the tension between rising demand and growing capacity is deep and substantial. Today, this tension concerns only a small fraction of the electromagnetic spectrum as smartphone and Wi-Fi communications predominantly use portions of the spectrum between 800 MHz and 2.5 GHz.

Two trends are likely to shift our historical approach of fixed spectrum allocation to much more nimble and dynamically adaptive management. The first is the rise of NUIs, which will necessarily embody data fusion and presentation from multiple sensors and devices that operate across differing communication bands and distances. From body and room area communications based on technologies such as Bluetooth; 60-GHz high-definition video transmissions; Wi-Fi, SuperWi-Fi (whitespace), and cellular communications across tens of meters to kilometers; and low-bandwidth but even longer-range communications for smart grid and intelligent transportation systems, multimodal communications will be the norm, rather than the exception, for creating and delivering integrated experiences.

Continuing growth in expectations for anywhere, anytime information access, together with multimodal interaction via speech, vision, and gesture, will accelerate the adoption and deployment of cognitive radio communication protocols and standards. Real-time negotiation for spectrum access based on multiple criteria (power levels, priorities among usage classes, payment or access method, and propagation characteristics) is likely to become much more common.¹⁵ In turn, this will necessitate new spectrum management policies and regulation, radio and protocol standards, and classes of licensed and unlicensed spectrum access providers.

Shifting privacy norms

The rapid growth of digital information and our increasingly large digital personas, defined by social network interactions, Web browsing, digital commerce, and location-specific services, all bring new concerns about digital privacy and information management. Most of our notions of privacy and security are rooted in concepts of physical location and person, yet digital information crosses organizational and jurisdictional boundaries rapidly and easily, moving from device to device and among cloud service datacenters. Likewise, most of our concepts of access are binary—data is encrypted and inaccessible or in the clear and available.

However, the realities of information access are far more nuanced and subtle, as anyone who has configured information-sharing specifications on a social network

knows. The notion of “friend” is a complex and shifting concept based on perception, recent behavior by the relevant parties, and mood. Equally importantly, what is considered acceptable behavior differs widely by region, culture, and generation.

All of this suggests that we must rethink our notions of privacy and security and the associated technical mechanisms for protecting data, processes, and individuals. Just as public-key cryptography empowered individuals and organizations to store, process, and transmit data securely, we need new protocols and standards that reflect the complex realities of today’s interactions. These are likely to be based on at least three information-sharing principles:

- *lifetime constraints*—limiting the interval during which data can be accessed;
- *transitivity bounds*—defining the degree of sharing beyond an initial recipient; and
- *claims-based access*—specifying the purposes for which the data can be used.

Using such a model, a mobile user might allow her location to be used for the next five seconds by the merchant operating the store in which she is standing, but only to provide directions to the shelf where the desired product is available. No other usage, retention, or sharing would be allowed. The challenge is in allowing such user control while making the specifications intuitive, simple, and non-invasive and while creating the conditions that allow new uses and economic models to grow and flourish.

LOOKING FORWARD: RESEARCH AND SOCIETAL IMPLICATIONS

The research challenges and opportunities raised by the technical aspects of Dennard scaling in the deep nanometer regime, the growth of the Internet of Things, cloud services, and their fusion via NUIs cross the entire spectrum of computer science and engineering. How do we design cloud services to manage and integrate a plethora of Internet-connected devices that appear as a continuous information fabric? Can cognitive radio provide the needed spectrum capacity for these data streams? What software and cryptographic security technologies can best protect our digital personas? Moreover, optimizing data-center energy and water efficiency while still allowing global deployment in areas subject to resource constraints is a major area of activity, as is managing resilience and service reliability.

Despite their diversity, most of these technical challenges have a common theme: they are systemic issues for which finding a solution requires insights and expertise from multiple subdisciplines in computing and communications. As such, their solution will ultimately require us to rethink elements of computing research and education,

emphasizing parallel and distributed computing as normal rather than being an exception, studying component and subsystem interactions in the large for systemic reliability and resilience, and considering human behavioral dynamics and interaction.

These technological trends also have disruptive social implications, as the rate of technology change increasingly challenges the way our social structures adapt and respond. Shifts that once took a generation now occur in a few years, with concomitant economic, social, and governmental disruptions. To address these problems effectively, we need to attract, educate, and employ a computing technology workforce that is broadly representative of the population. Instead of focusing on hardware and software, computing in the 21st century must encompass the deeply important applications of technology in its broadest sense.

This means examining the legal, social, and cultural expectations associated with personalized robotics and autonomous vehicles and how they might facilitate the independence, mobility, and safety of an aging population. It means considering how the predictive and analytic power of the Internet of Things and the cloud might sustain economic development via integration of renewable energy sources, smart grids, and intelligent transportation systems, while also reducing per capita energy consumption. It requires determining how truly personalized medicine, with data gleaned from a personal web of sensors, can reduce healthcare costs and improve quality of life without compromising personal privacy. It also means exploring how broadband access and multimedia presentation of education materials can deliver just-in-time training to a workforce now experiencing economic dislocation due to technological and business shifts.

These are exciting and challenging times. As the democratization of computing technology accelerates, computing in both the small and the large will create new opportunities. Computing will increasingly become a ubiquitous substrate for education, research, business, government, and social discourse, and it is our opportunity and responsibility to guide it in partnership with others. **C**

References

1. T. Hey, S. Tansley, and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009.
2. R.H. Dennard et al., "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, Oct. 1974, pp. 256-268.
3. G.E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, Apr. 1965, pp. 56-59.
4. J.F. Gantz and D. Reinsel, *The 2011 Digital Universe Study: Extracting Value from Chaos*, Int'l Data Corp., June 2011;

www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm.

5. K. Kinshumann et al., "Debugging in the (Very) Large: Ten Years of Implementation and Experience," *Comm. ACM*, July 2011, pp. 111-116.
6. J. Larus and G. Hunt, "The Singularity System," *Comm. ACM*, Aug. 2010, pp. 72-79.
7. C.P. Thacker et al., "Alto: A Personal Computer," *Computer Structures: Principles and Examples*, D.P. Siewiorek, C.G. Bell, and A. Newell, eds., McGraw-Hill, 1982, pp. 549-572.
8. K.H. Davies, R. Biddulph, and S. Balashek, "Automatic Speech Recognition of Spoken Digits," *J. Acoustical Soc. Am.*, Nov. 1952, pp. 637-642.
9. R. Szeliski, *Computer Vision: Algorithms and Applications*, Springer, 2010.
10. V. Pavlovic, R. Sharma, and T. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence*, July 1997, pp. 677-695.
11. J. Shotton, et al., "Real-Time Human Pose Recognition in Parts from Single Depth Images," Microsoft Research, 2011; <http://research.microsoft.com/pubs/145347/BodyPartRecognition.pdf>.
12. Y.J. Lee, C. Zitnick, and M. Cohen, "ShadowDraw: Real-Time User Guidance for Freehand Drawing," *Proc. SIGGRAPH, ACM*, 2011; <https://webpace.utexas.edu/y136631~ylee/shadowdraw/ShadowDrawSiggraph11.pdf>.
13. P. Bahl et al., "White Space Networking with Wi-Fi-Like Connectivity," *SIGCOMM Computer Comm. Rev.*, Oct. 2009; <http://ccr.sigcomm.org/online/files/p27.pdf>.

Daniel A. Reed is Microsoft's corporate vice president for technology policy. Before coming to Microsoft, he was Chancellor's Eminent Professor at the University of North Carolina at Chapel Hill and founding director of UNC's Renaissance Computing Institute. Prior to that, he was a Gutgsell Professor, head of the Department of Computer Science, and director of the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. Reed received a PhD in computer science from Purdue University. He is a Fellow of IEEE and ACM. Contact him at reed@microsoft.com.

Dennis B. Gannon is director of the Cloud Computing Research Engagement project in the technology policy group at Microsoft. Prior to joining Microsoft, Gannon was a professor of computer science in the School of Informatics and Computing at Indiana University, where he also served as science director of the Pervasive Technology Labs. He received a PhD in mathematics from the University of California, Davis, and a PhD in computer science from the University of Illinois at Urbana-Champaign. Contact him at dennis.gannon@microsoft.com.

James R. Larus is a director in the Microsoft Research eXtreme Computing group, where he currently works on programming models for cloud computing and custom hardware for large-scale computing. He received a PhD in computer science from the University of California, Berkeley. Larus is a Fellow of ACM and a member of IEEE. Contact him at larus@microsoft.com.