

Green Computing

Are you ready for a personal energy meter?

ANDY HOOPER INSISTS he's not a utopian, but his vision of the future of computing shares some resemblances with the dreams of science-fiction writers.

He foresees a not-too-distant time when the world's sources of computing power are concentrated in remote server warehouses strategically located near the sources of renewable energy that power them, such as wind and solar farms. And the usage of the power sources could shift across the globe, depending on where energy is most abundant.

"The system we now employ is hugely wasteful," says Hopper, a professor of computer technology at the University of Cambridge and head of its Computer Laboratory. "We lose energy by relying on the national grid. I propose a system that is more efficient, much less expensive, and that would have an immediate impact on the world's energy consumption. It's always cheaper to move data than energy."

Hopper is among the more conspicuous and outspoken pioneers in the green computing movement—a multifaceted, global effort to reduce energy consumption and promote sustainability. Proposed and existing strategies range from the practical to the fanciful, and include government regulations, industry initiatives, environmentally

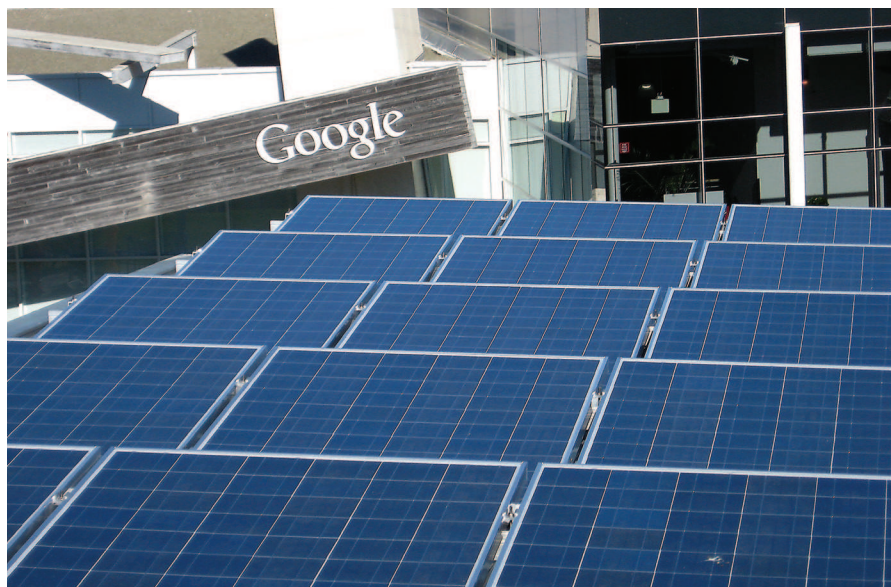
friendly computers made of recyclable materials, and Hopper's suggestion of a personal energy meter.

Much of the green computing movement's focus today is on data centers, which have been lambasted as "the SUVs of the tech world" for their enormous and wasteful consumption of electricity. The approximately 6,000 data centers in the United States, for instance, consumed roughly 61 billion kilowatt-hours (kWh) of energy in 2006, according to Lewis Curtis, a strategic infrastructure architect at Microsoft. The total cost of that energy, \$4.5 billion,

was more than the cost of electricity used by all the color televisions in the U.S. in 2006, Curtis says.

The Department of Energy (DOE) reports that data centers consumed 1.5% of all electricity in the U.S. in 2006, and their power demand is growing 12% a year. If data centers' present rate of consumption continues, Curtis warns, they will consume about 100 billion kWh of energy at an annual cost of \$7.4 billion by 2011.

The federal government wants data centers' energy consumption to be reduced by at least 10% by 2011. That translates into an energy savings equivalent to the electricity consumed by a million average U.S. households, according to Paul Sheathing, a spokesman for DOE's Office of Energy Efficiency and Renewable Energy.



"There's no simple path to green computing, but there are some low-hanging fruit," Curtis notes in "Green: The New Computing Coat of Arms?", a paper he co-authored with Joseph Williams, the CTO of WW Enterprise Sales at Microsoft. "You can spin the dial on some straightforward actions, such as orienting racks of servers in a data center to exhaust their heat in a uniform direction, thus reducing overall cooling costs.... A comprehensive plan for achieving green computing really does require an architectural approach."

David Wang, the data center architect for Teradata, has specialized in thermal management solutions for the Miamisburg, OH-based data warehousing company since 1996. "I've raised the issue [of green computing] because, for me, it's both a business question and an ethical question," Wang says. "Look at the basic fact, the one that has to be addressed: Power consumption at the server level has increased along with performance increase, and business needs have grown even faster."

More attention must be devoted to data centers' ever-increasing power density and heat removal, Wang says. "In the past, the sole focus was on IT equipment processing power and associated equipment spending. The infrastructure—power, cooling, data center space—was always assumed to be available and affordable," he says. "Now the infrastructure is becoming a limiting factor."

Microsoft, Google, and Yahoo are addressing the environmental concerns about their data centers' carbon footprint, the measure of the environmental impact of an individual or organization's lifestyle or operation, measured

Google uses customized evaporative cooling to significantly reduce its data centers' energy consumption.

in units of carbon dioxide produced.

In recent years, Microsoft and other companies have built data centers in central Washington to take advantage of the hydroelectric power produced by two dams in the region. The Microsoft facility, which consumes up to 27 megawatts of energy at any given time, is powered by hydroelectricity.

"This way, because we're so close to the source, we're not losing any energy and the energy we do use is pure and clean," says Francois Ajanta, Microsoft's director of environmental strategy.

Another Microsoft data center, located in Dublin, Ireland, is expected to become operational in 2009 and, thanks to Ireland's moderate climate, the 51,000-square-meter facility will be air cooled, making it 50% more energy-efficient than other comparably sized data centers.

Google "has committed to being carbon-neutral for 2007 and beyond," says Bill Weihl, Google's director of energy strategy. "Our carbon footprint is calculated globally and includes our direct fuel use, purchased electricity, and business travel—as well as estimates for employee commuting, construction,

and server manufacturing at our facilities around the world."

According to Google, its data centers use half the industry's average amount of power. Google attributes this improved energy usage to the cooling technologies, such as ultra-efficient evaporative cooling, that the company has customized for itself.

Yahoo's data centers also went carbon-neutral last year, in part because of its use of carbon offsets.

Government regulations and industry initiatives are also tackling data centers' energy usage. The U.S. Environmental Protection Agency (EPA), for instance, should have its phase-one version of Energy Star standards for servers ready by year's end. Eventually, the server rating will measure energy use at peak demand, but for the purpose of getting an Energy Star rating under way, the EPA will first release a Tier 1 standard, which will measure the efficiency of the server's power supply and its energy consumption while idle.

Meanwhile, a global consortium of computer companies, including AMD, Dell, IBM, Sun Microsystems, and VMware, organized The Green Grid in 2007, with the goal of improving energy efficiency in data centers and business computing systems. To achieve that goal, The Green Grid collaborates with individual companies, government agencies, and industry groups to provide recommendations on best practices, metrics, and technologies that will improve data centers' energy efficiency.

Earth-Friendly Computers

As with any evolving idea, people will need to think differently and more deeply when it comes to green comput-

Data Mining

Consumers' Invisible Profiles

Health and life insurance companies in the U.S. are increasingly using consumers' prescription drug data to determine what type of coverage, if any, to offer applicants, the *Washington Post* reports.

The insurance companies hire health information services companies—such as

Ingenix, which had \$1.3 billion in sales last year—to help create consumer profiles. The health information services companies mine the databases of prescription drug histories that are kept by pharmacy benefit managers (PBMs), which help insurers to process drug claims. (Ingenix even has its

own servers located in some PBM data centers.) The health information services companies also access patient databases held by clinical and pathological laboratories.

The health information services companies say that consumers have authorized the release of their records

and that their approach saves insurance companies money and time. Privacy advocates note that consumers do sign consent forms authorizing the release of data, but they have to if they want insurance, and that many people are unaware of the existence of health information services companies.

ing. It is not unusual, for instance, for companies to replace their older computers with new, more energy-efficient ones in an effort to become more earth-friendly.

This practice might not always be the most environmental solution, says Tera-data's Wang. "What I propose is that we look at the entire life cycle of a computer, the whole picture, from manufacturing through day-to-day operation," says Wang. "Every step consumes energy, and buying a new, more efficient computer may not always be the answer."

Some computer manufacturers are retooling their products from a life-cycle point of view and making the decision to buy a new, energy-efficient computer much easier. Dell is accelerating its programs to reduce hazardous substances in its computers, and its new OptiPlex desktops are 50% more energy-efficient than similar systems manufactured in 2005, thanks to more energy-efficient processors, new power management features, and other factors.

Likewise, Hewlett-Packard recently unveiled what it calls "the greenest computer ever"—the rp5700 desktop PC. The rp5700 exceeds U.S. Energy Star 4.0 standards, has an expected life of at least five years, and 90% of its materials are recyclable. The computer is easy to disassemble and meets the European Union's RoHS standards for the restriction of the use of certain hazardous substances in electrical and electronic equipment. Moreover, 25% of the rp5700's packaging materials are made of recycled material.

For the Future of the Planet

In an effort to ensure "computing can have a positive effect on our lives and the world," Hopper and Andrew Rice, an assistant director of research at the University of Cambridge's Computer Laboratory, have identified four principal goals in their paper "Computing for the Future of the Planet." The first goal is an optimal digital infrastructure in which computing's overall energy consumption is reduced and the efficient use of energy in the manufacture, operation, and disposal of computing devices is maximized.

The second goal is "to sense and optimize the world around us with reference to a global world model," which would "inform us about the energy con-

sumption and other effects of our activities on the natural environment."

The third goal is a new emphasis on predicting and responding to future events by modeling their behavior. According to Hopper and Rice, "The traditional role of computing as an execution platform for these models will continue to be important and must grow in performance to service both the increasing demands of higher-fidelity models and also to accommodate any new overheads incurred by correctness checking."

Lastly, Hopper and Rice are "interested in the possible benefit of digital alternatives to our physical activities," such as electronic versions of printed newspapers, music downloads rather than physical CDs, and online shopping as opposed to visiting stores and supermarkets. According to Hopper and Rice, "One might argue that a total shift from physical to digital seems unlikely in today's world but for future generations this concept might seem as obvious as email is to us today."

"People in the developing world," Hopper and Rice note, "often live in resource-impoveryed environments so a physical-to-digital paradigm shift has the potential to enable activities that were hitherto prohibitively expensive, and to support development whilst minimizing its impact. We seek to unlock methods of wealth creation in the virtual world."

Hopper and Rice also suggest the development of a personal energy meter that would measure a person's direct and indirect daily consumption, with individualized breakdowns of "the energy costs of travel, heating, water-usage and transportation of food [that] will help us target areas for reduction in our environmental footprint.... The data collected will not only provide useful information for analyzing consumption patterns but also has the potential to help individuals identify alternatives to their current activities."

"I think we've only just started to address the issue" of green computing, says Hopper. "It's just on the cusp of becoming important, and I think business, not academia, has led the way. They are driven by pragmatic concerns." ■

Patrick Kurp is a freelance science writer in Bellevue, WA.

Artificial Intelligence

Super-computer Defeats Human Go Pro

The new Dutch supercomputer Huygens, armed with the MoGo Titan program, defeated a human professional Go player with a 9-stones handicap. The victory appears to be the first-ever defeat of a high-level human Go player by a supercomputer in an official match.

Until recently, scientists were unable to create a computer program capable of beating even many amateur-level Go players. This state of affairs changed in 2006 when programmers Sylvain Gelly and Yizao Wang devised a revolutionary algorithm that has enabled the MoGo Titan program to attain new heights; since August 2006, MoGo Titan has been ranked number one on the 9x9 Computer Go Server.

Teamed up with the Huygens supercomputer, MoGo Titan achieved a noteworthy victory as its opponent, Kim Myungwan, is an 8 dan pro (the highest level is 9 dan) and a seasoned international competitor. In fact, the day before Myungwan's official match with Huygens and MoGo Titan, he soundly defeated the duo in three blitz games played with varying handicaps.

"The current result forecasts that before 2020 a computer program will defeat the best human Go player on a 19x19 Go board in a regular match under normal tournament conditions," says professor Jaap van den Herik of Maastricht University which, with INRIA France, co-developed MoGo Titan. "This is remarkable, since around 2000 it was generally believed that the game of Go was safe to any attack by a computer program. The 9-stones handicap victory casts severe doubts on this belief."

The Korean-born Myungwan appears to have taken the defeat well. Two days after his loss to MoGo Titan, he won the 2008 U.S. Open.

Chipping Away at Greenhouse Gases

Power-saving processor algorithms have the potential to create significant energy and cost savings.

THE INFORMATION TECHNOLOGY industry is in the vanguard of “going green.” Projects such as a \$100 million hydro-powered high-performance data center planned for Holyoke, MA, and green corporate entities such as Google Energy, the search giant’s new electrical power subsidiary, are high-profile examples of IT’s big moves into reducing the greenhouse gases caused by computers.

However, the true benefits of such projects are likely to be limited; most users in areas supplied by coal, oil, or natural gas-fired power plants would likely find it difficult to change to a fully sustainable supply source.

These market dynamics have not been lost on government research directors. Agencies such as the U.S. National Science Foundation (NSF) have begun encouraging just the sort of research into component-level power management that might bring significant energy savings and reduced climatic impact to end users everywhere without sacrificing computational performance.

In fact, the NSF has held two workshops in the newly emphasized science of power management, one in

The screenshot shows the Granola application interface. On the left, it displays several statistics with icons: a lightbulb for 'You'll save 234.6 kWh yearly' (enough to power 31 electric furnaces, a space heater, and 2 refrigerators for an hour), a dollar sign for 'You'll save 28.15 USD yearly' (enough for a monkey wrench to throw in the gears, 3 shirts from the thrift store, and a political bumper sticker), a leaf for 'You'll save 319.0 lbs CO2 yearly' (as much as a 500-mile flight, a tree, and 14 miles in a compact car), and a percentage sign for 'You've saved 45.1% CPU energy ...and you didn't even notice!'. At the bottom left, it says '151,738 trees' and 'You and the Granola community will offset 151,738 trees worth of CO2 this year.' with an illustration of trees. On the right, there is a large yellow padlock icon with a 'g' on it. Below the padlock, it says 'By logging into your Granola account, you can track your savings across systems through the Granola website. Click the lock above to access your savings information.' There is a login form with fields for 'Username' (with a hint '*****@acm.org') and 'Password' (with a hint '*****'). A 'Sign In' button is next to the password field. Below the login form, there are links for 'Forgot your password?' and 'Need an account?'. At the top right of the interface, there are social media icons for Facebook, Twitter, and a settings gear.

An intelligent power-management application, Granola uses predictive algorithms to dynamically manage frequency and voltage scaling in the chips of consumer PCs.

2009 and one in 2010. Krishna Kant, a program director in the Computer Systems Research (CSR) cluster at the NSF, says the power management project is part of the NSF’s larger Science, Engineering, and Education for Sustainability (SEES) investment area.

“There are some fundamental ques-

tions that haven’t been answered, and NSF funding might help answer them,” Kant says. “These have been lingering for quite some time. For instance, when you look at the question of how much energy or power you really need to get some computation done, there has been some research, but it tends

to be at a very, very abstract level to the extent it's not very useful."

Thermal Head Start

However abstract the state of some of the research into power management might be, basic computer science has given the IT industry a head start over other industries in addressing power issues. Whereas an auto manufacturer could continue to make gas-guzzling vehicles as long as a market supported such a strategy, two factors in particular have focused microprocessor designers' efforts on the imperatives of power efficiency.

One of the factors is the thermal limitations of microprocessors as each succeeding generation grew doubly powerful per unit size. The other is the proliferation of laptops and mobile computing devices, which demand advanced power management features to extend battery life. Kirk Cameron, associate professor of computer science at Virginia Polytechnic Institute, says this shift in product emphasis has given engineers working on power management theories more tools with which to work on the central processing unit (CPU); these chips are also installed on desktop machines and servers as chip manufacturers design one family for numerous platforms, based on overall market demand. Examples of these tools include application programming interfaces such as Intel's SpeedStep and AMD's PowerNow, which allow third-party software to dynamically raise or lower the frequency of cycles and the voltage surging through the

processor, depending on the computational load at any given time.

However, the default power management schemes supported by current operating systems, which allow users to specify either a high-performance or battery-maximizing mode on laptops, for instance, have numerous handicaps, including their static nature. The fact they need to be manually configured hampers their popularity.

Some power-management products, incubated by university researchers, are already available to dynamically manage power within a computer's CPU. Cameron is also the CEO of Miserware, a startup funded in part by an NSF Small Business Innovation Research Grant. Miserware produces intelligent power-management applications—called Granola for consumer PCs and Miserware ES for servers—that use predictive algorithms to dynamically manage frequency and voltage scaling. Company benchmarks claim that users can reduce power usage by 2%–18%, depending on the application in use; best savings are generated by scaling down power during low-intensity activities.

Granola was launched on Earth Day last year, and has 100,000 downloads. Cameron says the dynamic voltage and frequency scaling (DVFS) technology is very stable, available on most systems, and "kind of the low-hanging fruit" in power management.

Susanne Albers, professor of computer science at Humboldt University of Berlin, believes speed scaling will be a standard approach to power manage-

ment for some time. "I am confident that dynamic speed scaling is an approach with a long-term perspective," she says. "In standard office environments the technique is maybe not so important. However, data and computing centers, having high energy consumption, can greatly benefit from it."

Multicore Architectures

Ironically, although the DVFS technology is currently the most ubiquitous power management solution for processors, Cameron and other researchers say new fundamentals of computing architecture will mandate wholly different solutions sooner rather than later.

The onset of mass production of multicore processors, for example, is mandating that researchers begin practically anew in exploring speed scaling approaches.

"Generally speaking, there exists a good understanding of speed scaling in single processor systems, but there are still many challenging open questions in the area of multicore architectures," Albers notes.

"The new technologies bring new algorithmic issues," says Kirk Pruhs, professor of computer science at the University of Pittsburgh, and an organizer of both NSF workshops. For instance, if a heterogeneous-cored processor is programmed correctly, the utility of using frequency and voltage scaling at all might be moot—applications needing lower power can be sent to a slower core.

However, Pruhs says programming these will be "much more algorithmi-

ACM Awards News

2011 ACM Fellows Nominations

The ACM Fellow program was established by the ACM Council in June 1993 to recognize outstanding ACM members for technical, professional, and leadership contributions that advance the arts, sciences, and practices of information processing; promote the free interchange of ideas and information in the field; develop and maintain the integrity and competence of individuals in the field;

and advance the objectives of ACM.

Each candidate is evaluated as a whole individual and is expected to bring honor to the ACM. A candidate's accomplishments are expected to place him or her among the top 1% of ACM members. In general, two categories of accomplishments are considered: achievements related to information technology and outstanding

service to ACM or the larger computing community. A person selected as an ACM Fellow should be a role model and an inspiration to other members.

Nominations and endorsements must be submitted online no later than Sept. 1, 2011. For Fellows Guidelines, go to http://awards.acm.org/html/fellow_nom_guide.cfm/.

Nomination information organized by a principal

nominator should include excerpts from the candidate's current curriculum vitae, listing selected publications, patents, technical achievements, honors, and other awards; a description of the work of the nominee, drawing attention to the contributions which merit designation as Fellow; and supporting endorsements from five ACM members. For the list of 2010's ACM Fellows, see p. 25.

cally difficult for the operating system to manage, and the same thing happens in memories. The fact everything is changing means you have to go back and reexamine all the algorithmic issues that arise.”

In the case of power management in a parallel environment, Cameron says his research has shown that one cannot take the principles of Amdahl's Law for parallelization—which states that any parallelized program can only speed up at the percentage of a given task within that program not run serially—and get a correct assumption about power savings by simply taking into account the processors running a given application.

“In Amdahl's Law, you have one thing that changes, the number of processors,” Cameron says. “In our generalization, we ask what if you have two observable changes? You might think you could apply Amdahl's Law in two dimensions, but there are interactive effects between the two. In isolation, you could measure both of those using Amdahl's Law, but it turns out there is a third term, of the combined effects working in conjunction, and that gets missed if you apply them one at a time.”

Doing Nothing Well

In the long term, power management may borrow from sensor networks and embedded systems, which have extensively dealt with power constraints. Both David Culler, professor of computer science at the University of California, Berkeley, and Bernard Meyer-son, vice president of innovation at IBM, cite the disproportionately large power demands of processors doing little or no work as an area where great savings may be realized.

Culler says processor design might take a lesson from network sensor design in principle. Measuring performance during active processing “talk” time is misplaced, he says. Instead, efficiency must be introduced while awaiting instruction—“talk is cheap, listening is hard.”

Culler says theories behind effectively shutting down idle processors (“doing nothing well”) essentially fall into two basic camps that “hearken back to dark ages”—the principles following Token Ring or other time

In the long term, processor power management may borrow from sensor networks and embedded systems, which have extensively dealt with power constraints.

division multiplex technologies, or a Carrier Sense Multiple Access approach akin to Ethernet topology, in which nodes about to transmit can first “sense” whether or not a network is idle before proceeding.

He says this principle can apply to any scenario, be it a Wi-Fi network or a bus protocol on a motherboard. “Doing nothing well and being able to respond to asynchronous events anyway is the key to power proportionality, and can apply across the board,” says Culler.

Management From a Chip

Market demand for dynamically provisioned processors is still an unknown. Albers says processor-level power management is not particularly viewed as a critical issue among European users.

“Energy and environmental issues have always received considerable attention in Europe. However, the typical person is probably more concerned about energy consumption in his household and private car than about the consumption of his PC or laptop,” Albers observes.

IBM has placed a bet on combining chip-level energy allotment with the network architectures of homes and offices. The company has introduced fabricating technology for dedicated power management chips that control power usage while they communicate wirelessly in real time with systems used to monitor smart buildings, energy grids, and transportation systems. The main function of power-management chips is to optimize power usage

and serve as bridges so electricity can flow uninterrupted among systems and electronics that require varying levels of current.

Meyerson says that, while reducing battery usage on end user devices may be sexy, “that’s not the win for society. The win for society is when there’s an area of a building and the sensors over a period of time crawl through all the data of the occupancy of all the offices, and they autonomically adjust for the fact this is Paris in August—and in Paris in August people just aren’t showing up.”

IBM estimates the new technology can cut manufacturing costs by about 20% while allowing for the integration of numerous functions, resulting in one chip where previously three or four were needed. Meyerson says the technology can work for any appropriate algorithm researchers can come up with.

“Discovery algorithms that can look ahead and be predictive instead of reactive can be incredibly important,” he says. “What we are doing is ensuring that if they come up with a solution, there’s a way to execute it in a single chip, in a very efficient, synergistic way. It is a real footrace to stay ahead of the energy demands of society and IT.” ■

Further Reading

Albers, S.

Energy-efficient algorithms, *Communications of the ACM* 53, 5, May 2010.

Bansal, N., Kimbrel, T., and Pruhs, K.

Speed scaling to manage energy and temperature, *Journal of the ACM* 54, 1, March 2007.

Ge, R. and Cameron, K.W.

Power-aware speedup. *IEEE International Parallel and Distributed Processing Symposium*, Long Beach, CA, March 26–March 30, 2007.

Gupta, R., Irani, S., and Shukla, S.

Formal methods for dynamic power management. *Proceedings of the International Conference on Computer Aided Design*, San Jose, CA, Nov. 11–13, 2003.

Yao, F., Demers, A., and Shenker, S.

A scheduling model for reduced CPU energy. *Proceedings of the 36th IEEE Symposium on Foundations of Computer Science*, Milwaukee, WI, Oct. 23–25, 1995.

Gregory Goth is an Oakville, CT-based writer who specializes in science and technology.

© 2011 ACM 0001-0782/11/0200 \$10.00

DOI:10.1145/1721654.1721673

Prior work on power management reflects recurring themes that can be leveraged to make future systems more energy efficient.

BY PARTHASARATHY RANGANATHAN

Recipe for Efficiency: Principles of Power-Aware Computing

POWER AND ENERGY are key design considerations across a spectrum of computing solutions, from supercomputers and data centers to handheld phones and other mobile computers. A large body of work focuses on managing power and improving energy efficiency. While prior work is easily summarized in two words—“Avoid waste!”—the challenge is figuring out where and why waste happens and determining how to avoid it. In this article, I discuss how, at a general level, many inefficiencies, or waste, stem from the inherent way system architects address the complex trade-offs in the system-design process. I discuss common design practices that lead to power

inefficiencies in typical systems and provide an intuitive categorization of high-level approaches to addressing them. The goal is to provide practitioners—whether in systems, packaging, algorithms, user interfaces, or databases—a set of tools, or “recipes,” to systematically reason about and optimize power in their respective domains.

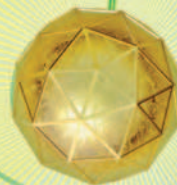
If you are a user of any kind of computing device, chances are you can share a personal anecdote about the importance of power management in helping control the electricity (energy) it consumes. On mobile devices, this translates directly into how long the battery lasts under typical usage. The battery is often the largest and heaviest component of the system, so improved battery life also enables smaller and lighter devices. Additionally, with the increasing convergence of functionality on a single mobile device (such as phone + mp3 player + camera + Web browser), battery life is a key constraint on its utility. Indeed, longer battery life is often the highest-ranked metric in user studies of requirements for future mobile devices, trumping even increased functionality and richer applications.

Power management is also important for tethered devices (connected to a power supply). The electricity consumption of computing equipment in a typical U.S. household runs to several hundred dollars per year. This cost is vastly multiplied in business enterprises. For example, servers in Google’s data centers have been estimated to consume millions of dollars

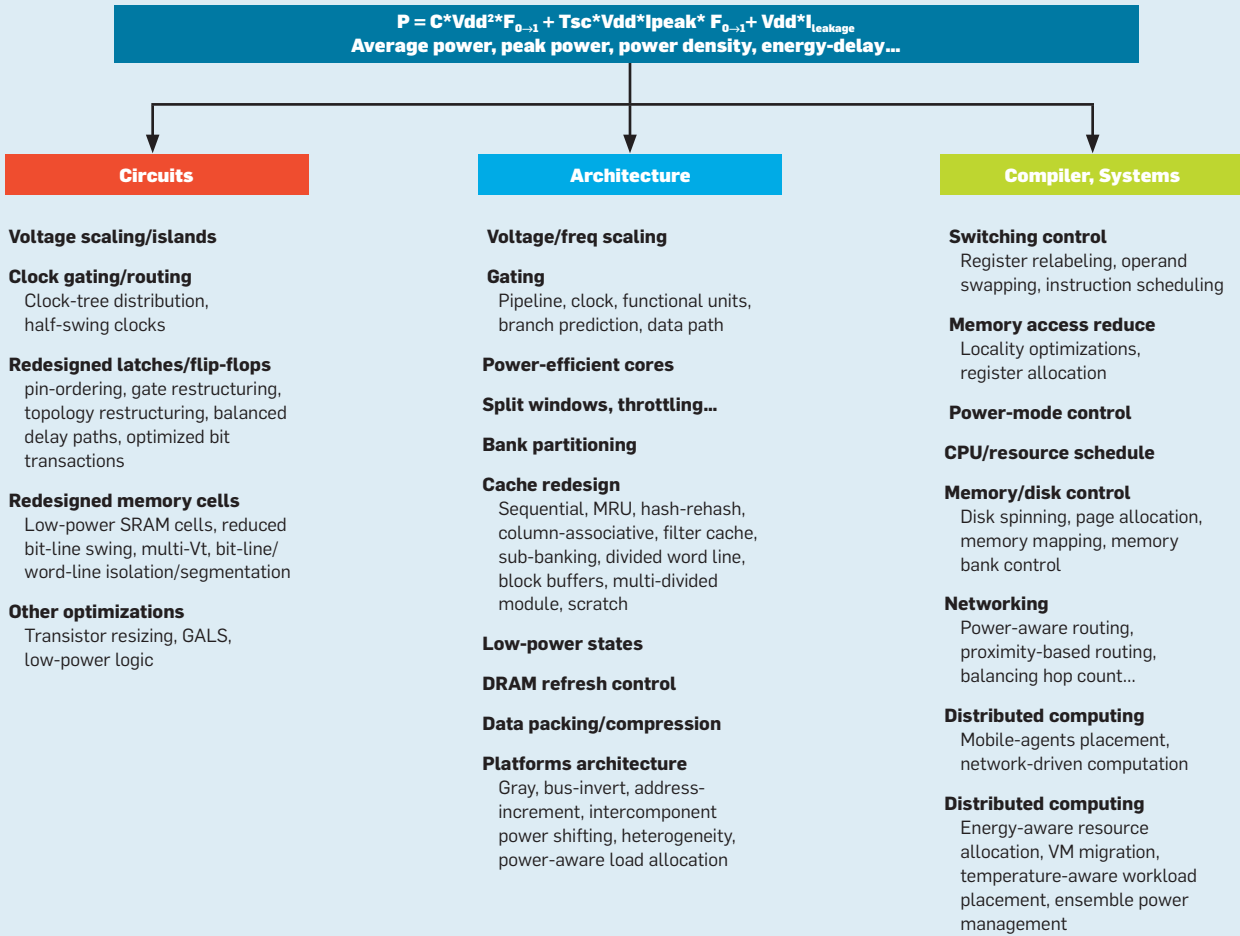
>> key insights

- **The energy efficiency of today’s systems can be improved by at least an order of magnitude.**
- **A holistic look at how systems use power and heat reveals new “recipes” to help optimize consumption and avoid wasting precious resources for a given task.**
- **Future power management will include nontraditional approaches, including crossing individual layers of design and spending more power to save power.**

ILLUSTRATION BY JEAN-FRANÇOIS PODEVIN



Overview of previous work on power management.



in electricity costs per year.¹⁰ IT analysis firm IDC (<http://www.idc.com/>) estimates the total worldwide spending on power management for enterprises was likely a staggering \$40 billion in 2009. Increased power consumption can also lead to increased complexity in the design of power supplies (and power distribution and backup units in larger systems) that also add costs.

Another challenge associated with power consumption in systems is the waste heat they generate; consequently, the term “power management” also includes the heat management in systems. Such heat is often a greater problem than the amount of electricity being consumed. To prevent the heat from affecting the user or the system’s electronics, systems require increasingly complex thermal packaging and heat-extraction solutions, adding more costs. For large systems like supercomputers and data centers, such costs often mean an additional

dollar spent on cooling for every dollar spent on electricity. This effect is captured in a metric called “power usage effectiveness,” or PUE,¹³ developed by the Green Grid, a global consortium of IT companies seeking to improve energy efficiency in data centers. Heat dissipation in systems also has implications for the compaction and density of computing systems, as in blade-server configurations.

Studies, most notably concerning servers and hard-disk failures, have shown that operating electronics at temperatures that exceed their operational range can lead to significant degradation of reliability; for example, the Uptime Institute, an industry organization that tracks data-center trends (<http://www.uptimeinstitute.org/>), has identified a 50% increased chance of server failure for each 10°C increase over 20°C¹⁵; similar statistics have also been shown over hard-disk lifetimes.^{1,4}


Finally, power management in computing systems has environmental implications. Computing equipment in the U.S. alone is estimated to consume more than 20 million gigajoules of energy per year, the equivalent of four-million tons of carbon-dioxide emissions into the atmosphere.¹⁰ Federal agencies have identified energy-consumption implications for air quality, national security, climate change, and electricity-grid reliability, motivating several initiatives worldwide from governmental agencies, including the Environmental Protection Agency in the U.S. (<http://www.epa.gov/>), Intelligent Energy Europe (ec.europa.eu/energy/intelligent/), Market Transformation Program in the U.K. (<http://efficient-products.defra.gov.uk/cms/market-transformation-programme/>), and Top Runner (http://www.eccj.or.jp/top_runner/index.html) in Japan, and from industry consortiums, including SPEC (<http://www.spec.org/>), Green-

Grid (<http://www.thegreengrid.org/>), and TPC (http://www.tpc.org/tpc_energy/default.asp) on improving energy efficiency, or minimizing the amount of energy consumed for a given task.


The importance of power management is only likely to increase in the future. On mobile devices, there is a widening gap between advances in battery capacity and anticipated increases in mobile-device functionality. New battery technologies (such as fuel cells) might address it, but designing more power-efficient systems will still be important. Energy-review data from the U.S. Department of Energy (<http://www.eia.doe.gov/>) points to steadily increasing costs for electricity. Indeed, for data centers, several reports indicate that costs associated with power and cooling could easily overtake hardware costs.^{2,14} Increased compaction (such as in future predicted blade servers) will increase power densities by an order of magnitude within the next decade, and the increased densities will start hitting the physical limits of practical air-cooled solutions. Research is ongoing in alternate cooling technologies (such as efficient liquid cooling), but it will still be important to be efficient about generating heat in the first place. All of this requires better power management.

How to Respond

Much prior work looked at power management and energy efficiency; the figure here outlines key illustrative solutions in the literature across different levels of the solution stack in process technology and circuits, architecture and platforms, and applications and systems design. A detailed discussion of the specific optimizations is not my intent here, and, indeed, several tutorial articles^{6,10,11} and conferences that focus solely on power, including the International Symposium on Low Power Electronics and Design (<http://www.islped.org/>) and the Workshop on Power Aware Computing and Systems (aka HotPower; <http://www.sigops.org/sosp/sosp09/hotpower.html>), provide good overviews of the state of the art in power management. This rich body of work examining power management and energy efficiency can be broadly categorized across different levels of



The goal is to provide practitioners a set of tools, or “recipes,” to systematically reason about and optimize power in their respective domains.



the solution stack (such as hardware and software), stages of the life cycle (such as design and runtime), components of the system (such as CPU, cache, memory, display, interconnect, peripherals, and distributed systems), target domains (such as mobile devices, wireless networks, and high-end servers), and metrics (such as battery life and worst-case power). Much prior work concerns electrical and computer systems engineering, with a relatively smaller amount in the core areas of computer science. The prior focus on power and energy challenges at the hardware and systems levels is natural and central, but, in the future, significant improvements in power and energy efficiency are likely to result from also rethinking algorithms and applications at higher levels of the solution stack. Indeed, discussions in the past few years on the future of power management focused this way.^{9,12}

In spite of the seemingly rich diversity of prior work on power management, at a high level, the common theme across all solutions is “Avoid wasted energy!” Where the solutions differ is in the identification and intuition needed for specific sources of inefficiency, along with the specific mechanisms and policies needed to target these inefficiencies. This observation raises interesting questions: What general recurring high-level trends lead to these inefficiencies at different levels of the system? And what common recurring high-level approaches are customized in the context of specific scenarios? The ability to answer supports the beginnings of a structure to think about power management in a more systematic manner and potentially identify opportunities for energy efficiency beyond traditional platform-centric domains.


Sources of Waste

It is easy to imagine that there is a certain minimum amount of electrical energy needed to perform a certain task and a corresponding minimum amount of heat that must be extracted to avoid thermal problems. For example, R.N. Mayo et al.⁸ performed simple experiments to measure the energy consumption of common mobile tasks (such as listening to music, making a phone call, sending email


and text messages, and browsing the Web) implemented on different devices (such as cellphones, MP3 players, laptops, and PCs) and observed two notable results: There is a significant difference in energy efficiency, often 10- to a hundredfold, across different systems performing the same task. And there are variations in the user experience across devices, but even when focused on duplicating the functionality of the best-performing system, these experiments showed it was impossible to do so at the same energy level on a different worse-performing system.

Why do some designs introduce additional inefficiencies over and above the actual energy required for a given task? My observation is that these inefficiencies are often introduced when the system design must reconcile complex trade-offs that are difficult to avoid. For example, systems are often designed for the most general case, most aggressive workload performance, and worst-case risk tolerance. Such designs can lead to resource overprovisioning to better handle transient peaks and offer redundancy in the case of failure. Moreover, individual components of a broader system are often designed by different teams (even by different vendors) without consideration for their use with one another. Individual functions of a system are also designed modularly, often without factoring their interactions with one another, adding further inefficiencies. Further, traditional designs focus primarily on system performance. This approach has sometimes led to resource-wasteful designs to extract small improvements in performance; with today's emphasis on energy costs, these small improvements are often overshadowed by the costs of power and heat extraction. Similarly, additional inefficiencies are introduced when the system design takes a narrow view of performance (vs. actual end-user requirements) or fails to address total cost of ownership, including design and operational costs.

General-purpose solutions. General-purpose systems often provide a better consumer experience; for example, most users prefer to carry a single converged mobile device rather than sev-



An insidious problem is when each layer of the stack makes worst-case assumptions about other layers in the stack, leading to compound inefficiencies.



eral separate devices (such as phone, camera, and MP3 player or GPS unit). Additionally, the exigencies of volume economics further motivate vendors to develop general-purpose systems; a product that sells in the millions of units is usually cheaper to make than, say, a product that sells in the hundreds of units.

By definition, general-purpose systems must be designed to provide good performance for a multitude of different applications. This requirement results in designers using the “union” of maximum requirements of all application classes. For example, a laptop that targets a DVD-playback application might incorporate a high-resolution display and powerful graphics processor. When the same laptop is used for another task (such as reading email), the high-power characteristics of the display and graphics processor might not be needed. However, when the laptop is designed for both workloads, most designs typically include a display with the characteristics of the most aggressive application use, in this case, a high-resolution display that plays DVD movies well. Lacking adequate design thought into how energy consumption might be adapted to different kinds of tasks, such an approach often leads to significant power inefficiencies. Another example is in the data center, where optimizing for both mission-critical and non-mission-critical servers in the same facility can lead to significant inefficiencies in terms of cooling costs. Similar conflicting optimizations occur when legacy solutions must be supported on newer systems.

Planning for peaks and growth. Most workloads go through different phases when they require different performance levels from the system. For example, several studies have reported that the average server utilization in live data centers can be low (often 10%–30%). Mobile systems have also been found to spend a significant fraction of their time in idle mode or using only a small fraction of their resources.

However, most benchmarks (the basis of system design) are typically structured to stress worst-case performance workloads irrespective of how the system is likely to be used in prac-

tice. Consequently, many systems are optimized for the peak-performance scenario. In the absence of designs that proportionally scale their energy with resource utilization, the result can be significant inefficiencies. For example, many power supplies are optimized for peak conversion efficiency at high loads. When these systems are operated at low loads, the efficiency of conversion can drop dramatically, leading to power inefficiencies.

Similar overprovisioning occurs when planning for the future. Most computing systems are designed for three-to-five-year depreciation cycles, and in the case of larger installations, like data centers, even longer. Systems must be designed to ensure that sufficient capacity is built in to meet incremental growth needs. On many systems, overprovisioning also leads to inefficiencies when the system is not operating at the resource-utilization capacities that account for future growth. For example, a data center with cooling provisioned for one megawatt of operational power, but operating at only 100 kilowatts of power consumption, is significantly more inefficient than a data center with cooling provisioned for, say, 150 kilowatts of operational power and

operating at 100 kilowatts of actual power consumption.

Design process structure. Current system-design approaches generally follow a structured process. System functionality is divided across multiple hardware components (CPU, chipset, memory, networking, and disk in a single system or different individual systems in a cluster) and software components (firmware, virtualization layer, operating systems, and applications). Even within a component (such as the networking stack), there are often multiple layers with well-defined abstractions and interfaces. Power management is usually implemented within these well-defined layers but often without consideration for the interaction across the layers. However, such modular designs or local optimizations might be suboptimal for global efficiency without communication across layers. An insidious problem is when each layer of the stack makes worst-case assumptions about other layers in the stack, leading to compound inefficiencies.

Information exchange across layers often enables better power optimization. For example, a power-management optimization at the physical layer of a wireless communication

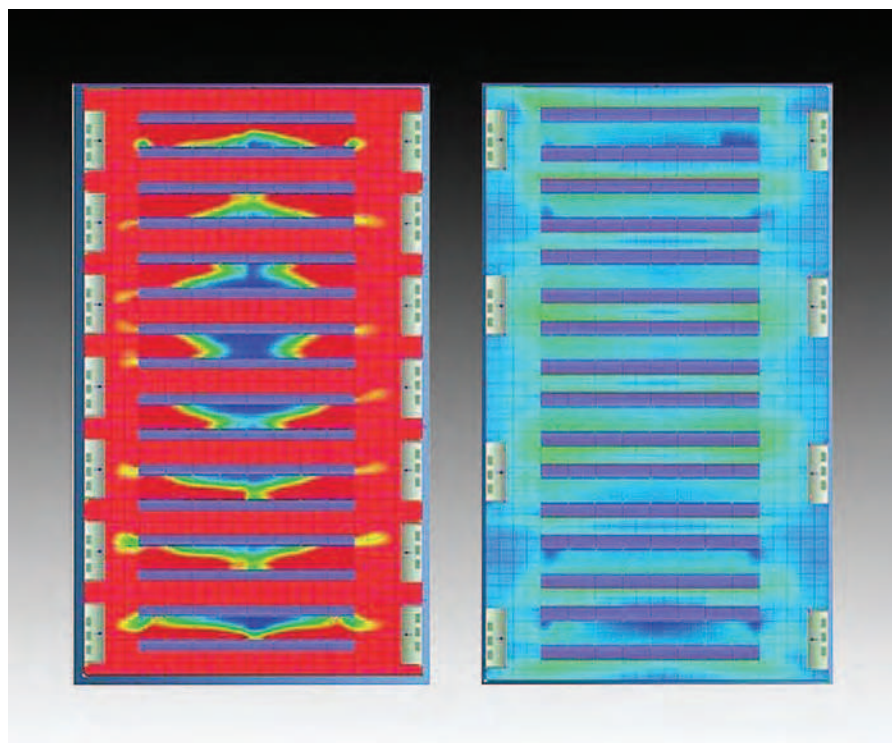
protocol that is aware of higher-level application activity can be more efficient than one that is oblivious to higher-level application activity. Similarly, a power-management solution that optimizes at an ensemble level (such as across different components in a system or different systems in a cluster) can be more efficient.

Similar problems exist at other boundaries of the system architecture. For example, the power management of servers is handled by the IT department, while the cooling infrastructure is often handled by a separate facilities department. This organizational structure can lead to inefficiencies as well. For example, a cooling solution that is aware of the nonuniformities in power consumption (and consequent heat generation) can be more efficient than a solution that is not.

Inefficiencies that result from layering can also be found at other places in the overall solution architecture. For example, in a classic client-server architecture, selectively exchanging information between the clients and servers has been shown to be beneficial for energy optimizations at both levels.

Tethered-system hangover. In this final design practice that leads to inefficiencies, the inefficiencies are mainly a reflection of the relentless drive to achieve higher performance, often following the assumption that there is no constraint on power, particularly by tethered systems (plugged into a power supply when in use) with no immediate consideration of battery life. For example, historically, many processor-architecture designs have included optimizations that achieved incremental performance improvements inconsistent with the amount of additional power consumed to implement the solutions. Similar trade-offs are seen in designs for high availability at the expense of energy (such as triple modular redundancy running three concurrent executions of the same task to ensure no possibility of downtime).

Additional examples include designs with user interfaces that identify the content of interest to the user; expending energy in these areas can be more energy efficient than designs that focus on metrics like refresh rate.



IBM uses thermal analysis to test and create “green” configurations of its iDataPlex system to deliver optimal energy efficiency, as shown on the right.


Similarly, designs that focus on energy delay may be significantly more energy efficient but with only a marginal difference in performance from pure performance-centric designs. In general, several significant power inefficiencies in today's systems stem from a design focus that does not sufficiently address total cost of ownership and ultimate end-user experience, but rather focuses disproportionately on one or more narrow metrics.

How to Reduce Waste


Once these inefficiencies are identified, the next step is to identify approaches to reduce them that fall into 10 broad categories:

Use a more power-efficient alternative. These approaches include replacing a system component with a more power-efficient alternative that performs the same task with less energy. For example, more energy-efficient nonvolatile memory can replace a disk drive, and optics can replace conventional networking. A more power-efficient alternative might sometimes involve adding the right hooks to enable the approaches discussed later. For example, replacing a display with a single backlight with an alternate display that provides more fine-grain control of power can, in turn, enable power optimizations that turn off unused portions of the display. Choosing a power-efficient alternative often involves other trade-offs, possibly due to costs or performance; otherwise, the design would have used the power-efficient option in the first place.

Create "energy proportionality" by scaling down energy for unused resources. These approaches involve turning off or dialing-down unused resources proportional to system usage, often called "energy proportionality"² or "energy scale-down."⁸ Automatically turning off unused resources requires algorithms that respond to the consequences of turning off or turning down a system (such as by understanding how long it takes to bring the system back on again). If a single component or system lacks the option to be scaled-down, the optimization is sometimes applied at the ensemble level; examples of ensemble-level scale-down include changing traffic routing to turn off unused switches



Decades ago, Nobel physicist Richard Feynman implied we should be able to achieve the computational power of a billion desktop-class processors in the power consumption of a single typical handheld device.



and virtual-machine consolidation to coalesce workloads into a smaller subset of systems in a data center.

Match work to power-efficient option. These approaches are complementary to the preceding approach—energy proportionality—but, rather than having the resources adapt when not fully utilized for a given task, they match tasks to the resources most appropriate to the size of the task. An example is the intelligent use of heterogeneity to improve power efficiency (such as scheduling for asymmetric and heterogeneous multicore processors). Matching work to resources implies there is a choice of resources for a given task. In cluster or multicore environments, the choice exists naturally, but other designs might need to explicitly introduce multiple operation modes with different power-performance trade-offs.

Piggyback or overlap energy events. These approaches seek to combine multiple tasks into a single energy event. For example, multiple reads coalescing on a single disk spin can reduce total disk energy. Prefetching data in predictable access streams or using a shared cache across multiple processes are other examples where such an approach saves energy. Disaggregating or decomposing system functionality into smaller subtasks can help increase the benefits from energy piggybacking by avoiding duplication of energy consumption for similar subtasks across different larger tasks.

Clarify and focus on required functionality. These approaches produce solutions specific to the actual constraints on the design without trying to be too general-purpose or future-proof. For example, special-purpose solutions (such as graphics processors) can be more energy-efficient for their intended workloads. Similarly, designs that seek to provide for future growth by adding modular building blocks can be more energy efficient compared to a single monolithic future-proof design.

Cross layers and broaden the scope of the solution space. Rather than having individual solutions address power management at a local level, focusing on the problem holistically is likely to achieve better efficiencies. Examples

where such an approach have been shown to be effective include scheduling across an ensemble of systems or system components and facilities-aware IT scheduling (such as temperature-aware workload placement). Exchanging information across multiple layers of the networking stack has also been shown to be beneficial for energy efficiency.

Trade off some other metric for energy. These approaches achieve better energy efficiency by marginally compromising some other aspect of desired functionality. An interesting example involves trading off fidelity in image rendering in DVD playback for extended player battery life. Also in this category are optimizations for improved energy delay where improvements in energy consumption significantly outweigh degradations in delay.

Trade off uncommon-case efficiency for common-case efficiency. These approaches seek to improve overall energy efficiency by explicitly allowing degradation in energy efficiency for rare cases and to improve energy efficiency in common cases. For example, a server power supply could be optimized for peak efficiency at normal light loads, even if it leads to degraded power efficiency at infrequent peak loads.

Spend someone else's power. These approaches take a more local view of energy efficiency but at the expense of the energy-efficiency of a different remote system. For example, a complex computation in a battery-constrained mobile device can be offloaded to a remote server in the “cloud,” potentially improving the energy efficiency of the mobile device. Approaches that scavenge energy from, say, excess heat or mechanical movement to improve overall energy efficiency also fall in this category.

Spend power to save power. A final category proactively performs tasks that address overall energy efficiency, even though these tasks may themselves consume additional energy. Examples include a garbage collector that periodically reduces the memory footprint to allow memory banks to be switched to lower-power states and a compression algorithm that enables the use of less energy for communication and storage.

The first five categories are well studied and found throughout existing power optimizations. The other five are less common but likely to be important in the future. Combinations are also possible.

Finally, irrespective of which approach is used to improve power efficiency, any solution must include three key architectural elements:

- ▶ Rich measurement and monitoring infrastructure;
- ▶ Accurate analysis tools and models that predict resource use, identify trends and causal relationships, and provide prescriptive feedback; and
- ▶ Control algorithms and policies that leverage the analysis to control power (and heat), ideally coordinated with one another.

From a design point of view, system support is needed at all levels—hardware, software, and application—to facilitate measurement, analysis, control, and cross-layer information sharing and coordination.

Looking Ahead

In spite of all this research and innovation, power management still has a long way to go. By way of illustration, several decades ago, Nobel physicist Richard Feynman estimated that, based on the physical limits on the power costs to information transfer,⁵ a staggering 10^{18} -bit operations per second can be achieved for one watt of power consumption. In terms easier to relate to, this implies we should be able to achieve the computational power of a billion desktop-class processors in the power consumption of a single typical handheld device. This is a data point on the theoretical physics of energy consumption, but the bound still points to the tremendous potential for improved energy efficiency in current systems. Furthermore, when going beyond energy consumption in the operation of computing devices to the energy consumption in the supply-and-demand side of the overall IT ecosystem (cradle-to-cradle³), the potential is enormous.

The energy efficiency of today's systems can be improved by at least an order of magnitude through systematic examination of their inherent inefficiencies and rethinking of their designs. In particular, in addition

to the large body of work in electrical and computer engineering, a new emerging science of power management can play a key role⁹ across the broader computer science community. I hope the discussions here—on the design practices that lead to common inefficiencies and the main solution approaches for addressing them—provide a starting framework toward systematically thinking about other new ideas in new domains that will help achieve the improvements. ■

References

1. Anderson, D., Dykes, J., and Riedel, E. More than an interface: SCSI vs. ATA. In *Proceedings of the Second Usenix Conference on File and Storage Technologies* (San Francisco, CA, Mar. 31–Apr. 2, 2003), 245–256.
2. Barroso, L.A. and Hölzle, U. The case for energy-proportional computing. *IEEE Computer* 40, 12 (Dec. 2007), 33–37.
3. Chandrakant, P. *Dematerializing the Ecosystem*. Keynote at the Sixth USENIX Conference on File and Storage Technologies (San Jose, CA, Feb. 26–29, 2008); <http://www.usenix.org/events/fast08/tech/patel.pdf>
4. Cole, G. *Estimating Drive Reliability in Desktop Computers and Consumer Electronics*. Tech. Paper TP-338.1. Seagate Technology, Nov. 2000.
5. Feynman, R. *Feynman Lectures on Computation*. Westview Press, 2000.
6. Irwin, M.J. and Vijaykrishnan, N. Low-power design: From soup to nuts. Tutorial at the International Symposium on Computer Architecture (Vancouver, B.C., June 10–14, 2000); <http://www.cse.psu.edu/research/mdl>
7. Lefurgy, C., Rajamani, K., Rawson, F., Felter, W., Kistler, M., and Keller, T.W. Energy management for commercial servers. *IEEE Computer* 36, 12 (Dec. 2003), 39–48.
8. Mayo, R.N. and Ranganathan, P. Energy consumption in mobile devices: Why future systems need requirements-aware energy scale-down. In *Proceedings of the Workshop on Power-Aware Computing Systems* (San Diego, CA, 2003), 26–40.
9. National Science Foundation. Workshop on the Science of Power Management (Arlington, VA, Apr. 9–10, 2009); <http://scipm.cs.vt.edu/>
10. Patel, C. and Ranganathan, P. Enterprise power and cooling: A chip-to-data-center perspective. In *Proceedings of Hot Chips 19* (Palo Alto, CA, Aug. 20, 2007); <http://www.hotchips.org/archives/hc19/>
11. Rajamani, K., Lefurgy, C., Ghiasi, S., Rubio, J.C., Hanson, H., and Keller, T. Power management for computer systems and datacenters. In *Proceedings of the 13th International Symposium on Low-Power Electronics and Design* (Bangalore, Aug. 11–13, 2008); <http://www.islped.org/X2008/Rajamani.pdf>
12. Ranganathan, P. (moderator). Power Management from Cores to Data Centers: Where Are We Going to Get the Next 10X? Panel at International Symposium on Low-Power Electronic Devices (Bangalore, 2008); <http://www.islped.org/X2008/>
13. Rawson, A., Pfeleuger, J., and Cader, T. (C. Belady, Ed.). *The Green Grid Data Center Power Efficiency Metrics: Power Usage Effectiveness and DCIE*. The Green Grid, 2007; www.thegreengrid.org
14. Shankland, S. Power could cost more than servers, Google warns. *CNET News* (Dec. 9, 2005); http://news.cnet.com/Power-could-cost-more-than-servers,-Google-warns/2100-1010_3-5988090.html
15. Sullivan, R.F. *Alternating Cold and Hot Aisles Provides More Reliable Cooling for Server Farms*. White paper, Uptime Institute, 2000; <http://www.dataclean.com/pdf/AlternColdnew.pdf>

Parthasarathy Ranganathan (Partha.Ranganathan@hp.com) is a distinguished technologist in Hewlett-Packard Labs, Palo Alto, CA.

SUSTAINABILITY IMPLICATIONS OF ORGANIC USER INTERFACE TECHNOLOGIES: AN INKY PROBLEM

BY ELI BLEVIS

The moment you decide sustainability is an issue with respect to interaction design and the design of interactive devices is the moment you realize how complex the business of deciding what to actually do about it is. It is not just a simple matter of calculating the energy and environmental costs of manufacturing, use, salvage, and disposal of one technology over another.

For example, it was long ago claimed that computing technologies would create a paperless office—a claim that is not yet in sight. Many people print things rather than read on screen—they like to hold paper in their hands and mark things up. Ever since I acquired a portrait mode capable LCD monitor, I have mostly stopped printing things. I can now read and write an entire page of text on my 1200x1600 pixel screen at 140% the size it would be if I printed it. As a result, I almost never print anything anymore. The environmental costs of the energy used to power my display must be weighed against the costs of printing the page when I am just reading, assuming that I would actually power-off my display when I am reading what has been printed. Furthermore, the environmental cost of production of the portrait mode display and the environmental costs of the premature obsolescence and disposal of the display I had before this one are also part of the equation.

Environmental costs are not very static—increasing demands for a technology can drive down some such environmental costs while increasing some others. Nonetheless, Organic UI technologies, such as digital paper or flexible displays and E-Ink technologies offer promising potentials for the development of sustainable practices in interaction design. Each of these potentials has dangers of inducing unsustainable behaviors as well.

One potential is due to an advantage of paper display technology itself. No energy is used when reading an E-Ink display owing to the bistability of the material—that is, digital paper preserves its state each time it is updated without the need for additional power. From the perspective of environmental sustainability, this seems to be a more important feature than the issue of the present environmental cost of making “a sheet of” digital paper, since such costs will change dramatically with improvements

in the technology and with production on a larger scale.

A second potential is related to the concept of books as durable objects. When it becomes possible to create a book using the new digital paper that can be turned into any book by means of an electronic update, the potential for a more sustainable medium presents itself—one that does not require the cutting of trees. But, the durability of the digital paper book can only match that of the ordinary notion of a book if the other attributes that make ordinary books enduring objects in general are also matched or even exceeded.

A third potential is based on the possibilities for making displays that are more portable, cheaper, smaller, and more pervasive. From a sustainability point of view, pervasive, small, inexpensive displays may be an advantage to the degree that they build an infrastructure of modularity. If upgrading a display on an interactive device such as a cell phone, PDA, MP3/video player, or laptop becomes as viable as upgrading the storage capacity of a device by substituting a memory card such as an SD card, this could have the effect of making digital artifice last longer. On the other hand, if the possibility of making more portable, cheaper, smaller, and more pervasive displays ends up driving a practice of even more disposability and premature disposal due to frequent obsolescence with respect to display devices—for example, on product packaging—the consequences could be devastating from an environmental point of view. Even if the substrates are made of recyclable materials, recycling is not as environmentally sustainable as reuse. And, if the substrates are not made of recyclable or biodegradable materials, the effects on the e-waste stream may possibly augment the toxicity of the present-day e-waste stream [3]. In any event, the negative social impacts of adding to the e-waste stream and even of certain recycling practices are also a global sustainability issue [1, 2].

For digital paper to be better than ordinary paper

from a user experience point of view, it will need to properly address at least these four interactivity issues: resolution—the quality of the text will need to be as good or better than paper; control—the use of digital paper and labels will need to be as easy and straightforward in use as ordinary paper and labels; portability—digital paper will need to be as portable or more portable than ordinary paper at the same resolutions; authenticity—the experience of using these displays will need to be as aesthetically authentic and tangible as holding a physical



piece of paper. If these user experience concerns can be adequately addressed together with some of the other concerns described here, the potentials of organic display technologies to enable choices for a sustainable future can be realized. ■

REFERENCES

1. Iles, A. Mapping environmental justice in technology flows: Computer waste impacts in Asia. *Global Environmental Politics* 4, 4 (Nov. 2004); www.mitpressjournals.org/doi/pdfplus/10.1162/glep.2004.4.4.76.
2. Schmidt, C. Unfair trade e-waste in Africa. *Environmental Health Perspectives* 114, 4 (2006): A232–A235; www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1440802.
3. Townsend, G. et al. *RCRA Toxicity Characterization of Computer CPUs and Other Discarded Electronic Devices*. Department of Environmental Engineering Sciences, University of Florida, 2004.

ELI BLEVIS (eblevis@indiana.edu) is an assistant professor of informatics at Indiana University, Bloomington.

© 2008 ACM 0001-0782/08/0600 \$5.00

DOI: 10.1145/1349026.1349038

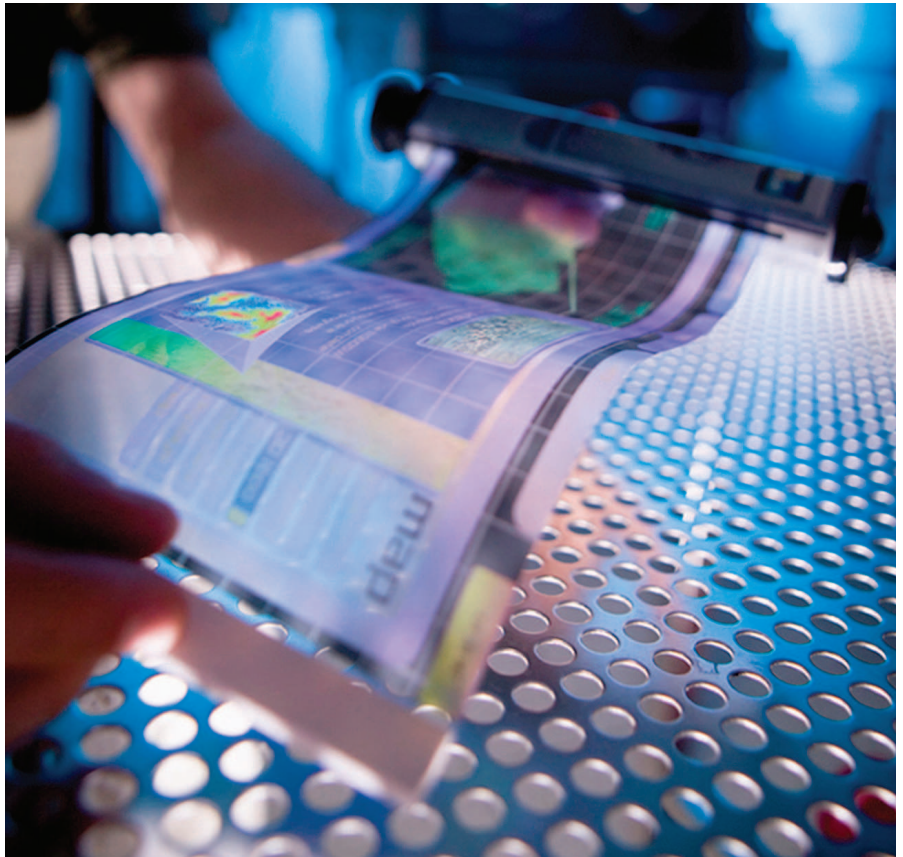
The Promise of Flexible Displays

New screen materials could lead to portable devices that are anything but rectangular, flat, and unbendable.

EVER SINCE SCIENTISTS first hacked oscilloscopes in the 1950s, computer displays have been heavy, fragile things. Even in a super-thin laptop, the quarter-kilogram screen is one of its heaviest parts—and the case required to protect it adds further bulk and weight. The screen also defines the form factor of most electronic devices as it's a single component that's rectangular, flat, and unbendable.

Now several developments portend an industrywide change to displays that are lighter, tougher, and more flexible than the sheet glass of a traditional liquid-crystal display (LCD). These displays are just beginning to appear on the market, and are produced with methods that may also improve energy efficiency, lower production costs, and allow for display shapes and sizes currently impossible to achieve.

Already these displays have started to appear in a variety of applications. The 75th anniversary issue of *Esquire* magazine gave them wide exposure by featuring e-paper that blinked “The future is now” on its front cover. Other uses already in the market include keyboards with auto-changing layouts, low-power shelf tags and point-of-sale ads, and display windows on credit-card-size smart cards. But designers are also rethinking product design around these new screens' possibilities. Carl Taussig, director of Advanced Display Research at Hewlett-Packard, notes the Dutch company Polymer Vision has demonstrated a cell phone with a roll-out display, and suggests other possible form factors. “You might have a display that you keep folded up like a piece of paper,” Taussig says. “You might open it halfway and use it that way, or you might open it all the way. You might have it partitioned with a keyboard on one part and a screen on the other.”



A flexible electronic display is rolled out at Arizona State University's (ASU's) Flexible Display Center. The unbreakable displays were developed by ASU and Hewlett-Packard Labs.

Front Plane, Back Plane

These new display technologies comprise two parts: a front plane that contains the imaging component, and a back plane that controls which pixels are on. For a display to be fully flexible, both parts must have that characteristic. While dozens of companies are competing to provide back planes, two front plane technologies have taken the lead.

One is electronic paper, or e-paper, like that used in the Amazon Kindle, Barnes & Noble Nook, and Sony Reader Digital Book. The producer of the e-paper in all three devices is the Massachusetts-based E Ink Corporation. Those

devices currently pair the flexible display layer with an inflexible glass back plane. E-paper is a reflective display; its brightness comes from ambient light hitting the display's face, as on a sheet of paper. Pixels in E Ink's e-paper are electrically dipolar units colored white on one side and black on the other, floating in thick oil. Their refresh rate is currently quite low, at about four frames per second. E-paper is bistable, retaining its image until it receives the next signal. As a result it consumes energy only when transitioning from one image to the next.

The other front plane technology is organic light-emitting diodes (OLEDs),

which emit rather than reflect light. OLED displays consume more power than e-paper, but are brighter and have faster refresh rates. They have a higher lumen-per-watt rating—that is, they use less power for comparable brightness—than comparable LCDs because the latter cause liquid crystals to block photons coming from a backlight. On the other hand, OLEDs actually emit light themselves.

Nick Colaneri, director of the Flexible Display Center at Arizona State University, believes OLED displays may eventually consume as little as 1/10th the power of LCDs, even as they deliver “a kind of intangibly superior image quality. Put side-by-side, OLED displays are often seen as strikingly better looking.” That makes them a likely challenger in the \$100 billion LCD industry, although OLED displays are extra difficult to ruggedize because they’re far more sensitive to oxygen and moisture. (E-paper is more likely to be used in applications where power and reliability trump brightness and contrast.)

Some manufacturers have proposed trying to capture both markets with hybrid transfective versions based on e-paper, similar to the Pixel Qi LCD screen on One Laptop Per Child’s XO computer. That screen includes both a backlight for low-light situations and a reflective layer for easy reading when the sun is out. Transfective displays with e-paper or OLED front planes are not currently available, but they’re possible because the back plane could be made from reflective or transparent materials.

With their similarity to LCDs, the potential rewards for OLED displays are enormous. At the same time, e-paper is defining new market segments, notably in portable electronics and large-scale signage. According to Jennifer Colegrove, vice president of emerging display technology at DisplaySearch, flexible plastic displays will grow at an annual rate of nearly 60%, surpassing \$8 billion in sales in 2018 from its current level of about \$300 million. Colegrove believes that although flexible displays will have fast growth in the next several years, they won’t become truly mainstream before 2018, both because of technical problems and the task of making them

Flexible displays will have fast growth during the next several years, but won’t become mainstream before 2018, says Jennifer Colegrove.

cost competitive with traditional glass-based displays.

Beyond Rectangular Thinking

Commercial products are available for both e-paper and OLED displays, including e-readers and advertising displays. Today, e-paper claims far greater market pull than OLED, largely because of e-readers. But Sriram Peruvemba, E Ink’s chief marketing officer, points to other uses for e-paper that weren’t possible with previous display technologies. “Most requests I get from design engineers are for rectangular displays, because that’s all they were able to get before,” says Peruvemba. “I tell them, ‘Think outside the rectangle.’ With our flexible display, you can cut it in any shape, so it’s only limited by imagination. Right now the device has to accommodate the display, but in the future the display will accommodate the device.”

Shape isn’t the only feature available to this new breed of displays but unavailable to LCDs. Because an LCD’s transparent electrodes must remain a set distance apart, they can’t be flexible, and are therefore made of thick, rigid glass. Transporting large sheets of glass is difficult and expensive, so LCDs are generally limited to a few square feet in size. But flexible displays can be made on material that’s stored and shipped in a roll, much like a web-fed paper press, enabling display sizes of a few meters wide by hundreds of meters long.

But with that flexibility are some practical problems. “How do you handle flexible substrates for TFT [thin-

In Memoriam

Philippe Flajolet Dies at 62

Soon after Philippe Flajolet passed away from a serious illness on March 22, tributes started appearing in an online book of tribute at INRIA from colleagues, former students, and the legions of computer scientists who were influenced by his contributions to the study of algorithms.

Those who knew Flajolet, an INRIA research director, best remember him as more than just an important theorist, however; he was a proudly independent researcher, as well as a gifted raconteur and free spirit who loved to play practical jokes on friends using complex mathematical formulae.

While working at INRIA, Flajolet earned the nickname “Algorithmix,” a nod to the popular Asterix books. As Richard J. Lipton noted in a blog post after Flajolet’s death, the nickname could hardly have fit more perfectly: “More than his development of particular algorithms, one can credit him much toward the development of *algorithmics* as a professional discipline.”

From his perch at INRIA, Flajolet devoted most of his career to studying the computational complexity of algorithms, the theory of average-case complexity, his transform-based asymptotic analysis, and the symbolic method, a novel approach to deriving the properties of combinatorial objects.

“The contribution of Philippe to the research on algorithms was essentially analytical,” says Micha Hofri, a computer science professor at Worcester Polytechnic Institute, “and even his algorithmic innovations, such as approximate counting of elements in multisets, came as the result of a mathematical insight.”

Hofri recalls meeting Flajolet at INRIA in the 1980s, when Flajolet wagered a bottle of champagne over whether one of Hofri’s analyses had already been completed by another researcher. Hofri lost the bet, but made a lifelong friend. “It was a good bottle,” he recalls.

—Alex Wright

film transistor] manufacturing?” Colegrove of DisplaySearch has asked. “Glass is easy: It’s a rigid sheet, and you just put layers on it. But with flexible displays it can be hard to register for the very accurate positioning you need. Also, some plastics deform when you heat them up, which can affect manufacturing. Therefore, several new processes have been invented.”

To address registration problems, some manufacturers have championed a roll-to-roll process that allows them to print electronics with greater precision than traditional photolithography allows. Photolithography is an “exposure” system: Light affects chemistry that etches electronics, similar to how an old-fashioned photo enlarger affects an image on paper. By comparison, the roll-to-roll process places electronics directly on a substrate and is more similar to a modern inkjet printer.

Hewlett-Packard’s explorations with unusual form factors led it to close a contract with the U.S. Army for wrist-mounted devices that would make critical information more easily available to soldiers in the field. As designed, the devices are only about 1.5mm thick and feature a screen that curves halfway around the wrist, the other half being a set of flexible photovoltaic cells to provide power. Taussig expects to deliver prototypes to the Army by this autumn, and sees viability for such devices well beyond its first customer.

“We imagine a lot of use cases,” he says. “[A flexible plastic screen] could be built into the cuff of a delivery driver’s uniform and include mapping information and the like. Or maybe it’s

for a mechanic or technician who’s assembling jet aircraft, and won’t have to go back to a table to see a schematic. Or a health-care technician or nurse could keep track of medication and patients and such. It’s very hands free. If it gets smashed into a door it’s not going to break. It’s very low power, and it’s light.”

Hewlett-Packard’s devices will be built with E Ink’s e-paper displays on a plastic substrate. But plastic wouldn’t work for OLED displays as its permeability would doom the sensitive OLEDs to a short life. For those applications, Corning Inc. is developing a kind of flexible glass by extending its proprietary fusion process, which it already uses to create thin glass substrates such as the Eagle XG line. While glass is necessarily heavier and more sensitive to certain stresses, Corning flexible glass commercial program manager Jill VanDewoestine believes the industry will respond well to this development because of glass’ distinctive qualities.

“Glass is really the standard,” VanDewoestine says. “Its surface is very smooth; it’s stable; high process temperatures are possible; it’s transparent; and it’s an excellent barrier against oxygen and moisture. So people really want to work with flexible glass—it gives them the handleability of roll-to-roll processes with the characteristics of glass that make high-quality electronics possible.” The company is currently developing flexible glass that’s 1/10th of a millimeter thick and can be wrapped around a core with a three-inch radius.

But whether incorporating glass,

plastic, or some other material, flexible display technologies represent the leading edge in changing the shape of computing devices. Consider the laptop computer, which in essence comprises a screen, keyboard, and processor. Flexible rubber keyboards are already available at big-box stores, and several manufacturers make “virtual keyboards” that project laser images of the keys onto any flat surface. Flexible plastic electronics are in active development, with the Russian corporation Rusnano recently announcing development of a \$700 million factory to produce such electronics. The age of the Dick Tracy wrist computer is already upon us. With flexible displays leading the way, there’s no limit to the shapes of computers to come. **C**

Further Reading

Chung, I.-J. and Kang, I. Flexible display technology—opportunity and challenges to new business application, *Molecular Crystals and Liquid Crystals* 507, Sept. 2009.

FlexTech Alliance for Displays & Flexible Printed Electronics
<http://www.flextech.org/>

Forge, S., Blackman, C., and Lindmark, S. Plastic promises: the disruptive potential of OLEDs and e-paper for the European display industry, *Foresight* 11, 3, 2009.

Jeong, J.K. The status and perspectives of metal oxide thin-film transistors for active matrix flexible displays, *Semiconductor Science and Technology* 26, 3, 2011.

Tom Geller is an Oberlin, OH-based science, technology, and business writer.

© 2011 ACM 0001-0782/11/06 \$10.00

Milestones

American Academy Announces the Class of 2011

The American Academy of Arts and Sciences elected 212 new members in its Class of 2011, including nine computer scientists. As members of one of the most prestigious honorary societies and a center for independent policy research, the Class of 2011 includes some of the world’s most accomplished leaders in academia, business, public affairs, the humanities,

and the arts. Members contribute to studies of science and technology policy, global security, social policy and American institutions, the humanities, and education.

Newly elected members in the computer sciences are Edmund M. Clarke, Carnegie Mellon University; Edward W. Felten, Princeton University; Eric Horvitz, Microsoft Research;

Michael I. Jordan, University of California, Berkeley; Shree K. Nayar, Columbia University; Patricia Griffiths Selinger, IBM Almaden Research Center; Peter Williston Shor, Massachusetts Institute of Technology; and Avi Wigderson, Institute for Advanced Study. Leah H. Jamieson, Purdue University, was inducted in the section of engineering sciences

and technologies.

“It is a privilege to honor these men and women for their extraordinary individual accomplishments,” said Leslie Berlowitz, American Academy president. “The knowledge and expertise of our members give the Academy a unique capacity—and responsibility—to provide practical policy solutions to the pressing challenges of the day.”