Implications of Historical Trends in the Electrical Efficiency of Computing

Jonathan G. Koomey
Stanford University

Stephen Berard
Microsoft

Marla Sanchez
Carnegie Mellon University

Henry Wong
Intel

The electrical efficiency of computation has doubled roughly every year and a half for more than six decades, a pace of change comparable to that for computer performance and electrical efficiency in the microprocessor era. These efficiency improvements enabled the creation of laptops, smart phones, wireless sensors, and other mobile computing devices, with many more such innovations yet to come.

Valentine’s day 1946 was a pivotal date in human history. It was on that day that the US War Department formally announced the existence of the Electronic Numerical Integrator and Computer (ENIAC). ENIAC’s computational engine had no moving parts and used electrical pulses for its logical operations. Earlier computing devices relied on mechanical relays and possessed computational speeds three orders of magnitude slower than ENIAC.

Moving electrons is inherently faster than moving atoms, and shifting to electronic digital computing began a march toward ever-greater and cheaper computational power that continues even to this day. These trends proceed at easily measurable, remarkably predictable, and unusually rapid rates. For example, the number of transistors on a chip has doubled more or less every two years for decades, a trend that is popularly (but often imprecisely) encapsulated as Moore’s law (see Figure 1).

Moore’s law has been modified over the years in several different ways, as previous research has established. The trend, as Moore initially defined it, relates to the minimum component costs at current levels of technology. All other things being equal, the cost per component decreases as more components are added to a chip, but because of defects, the yield of chips goes down with increasing complexity. As semiconductor technology improves, the cost curve shifts down, making increased component densities cheaper (see Figure 1 in Moore’s 1965 paper).

In 1975, Moore modified his observation to a doubling of complexity every two years.
which reflected a change in the economics and technology of chip production at that
time as well as a change in his conceptualization of the law. That rate of increase in chip
complexity has held for more than three decades, which is a reflection mainly of the un-
derlying characteristics of semiconductor manufacturing during that period. As Ethan
Mollick explained, Moore’s law is in some sense a self-fulfilling prophecy—the industry’s
engineers have used Moore’s law as a benchmark to which they calibrated their rate
of innovation. This result is partly driven by business dynamics, as David E. Liddle
points out: “Moore’s law expresses that rate of semiconductor process improvement
which transfers the maximum profit from the computer industry to the semiconductor
industry.”

The striking predictive power of Moore’s law has prompted many to draw links be-
tween chip complexity and other aspects of computer systems. One example is the popu-
lar summary of Moore’s law (“computing performance doubles every 18 months”),
which correlates well to trends in personal computer systems to date but is a statement
that Moore never made. Another is “Moore’s law for power,” coined by Wu-chun Feng to
describe changes in the electricity used by computing nodes in supercomputer installa-
tions during a period of rapid growth in power use for servers (“power consumption
of compute nodes doubles every 18 months”).

This article describes the implications of the relationship between the processing
power of computers (which in the microprocessor era has been driven by Moore’s
law) and the electricity required to deliver that performance. Over the past 65 years,
the steps taken to improve computing performance also invariably increased the electrical
efficiency of computing, whether the logical gates consisted of vacuum tubes and diodes,
discrete transistors, or microprocessors. The most important historical effect of this rela-
tionship has been to enable the creation of mobile computing devices such as laptop
computers. If these trends continue, they will have important implications for more
widespread future use of mobile computing, sensors, and controls.

Methods for Deriving Trends
Analyze long-term trends is a tricky busi-
ness. Ideally we’d have performance and
energy use data for all types of computers in
all applications since 1946. In practice, such
data do not exist, so we compiled available
data to piece together the long-term trends
on the electrical efficiency of computation.

To estimate computations per kilowatt-
hour, we focused on the full-load computa-
tional capacity and the direct active electrical
power for each machine, dividing the num-
ber of computations possible per hour at
full computational load by the number of
kilowatt-hours consumed over that same
hour. This metric says nothing about the
power used by computers when they are
idle or running at less than full load, but it
is a well-defined measure of the efficiency
of this technology, and it helps show how
the technology has changed over time.

Measuring computing performance has always
been controversial, and this article will
not settle those issues. In 2007, William D.
Nordhaus published a sophisticated and
comprehensive historical analysis of com-
puting performance over time, and that’s
the source for performance data on which
we relied most heavily. We combined those
data with measured data on the power use
of each computer when operating at full
load to calculate computations per kilowatt-
hour. (More details on the data and methods
are available in our Web Extra appendix, see
http://doi.ieeecomputersociety.org/10.1109/
MAHC.2010.28).

Historical Trend Results
Figure 2 shows performance per computer
for all the computers included in Nordhaus’
analysis from 1946 onward. We also added
the 40 additional machines from this anal-
ysis for which measured power and perfor-
ance were available. The figure does not
include performance estimates for recent
large-scale supercomputers (for example,
those at www.top500.org), but it does in-
clude measurements for server models that
Computations per kilowatt-hour doubled every 1.57 years over the entire analysis period, a rate of improvement only slightly slower than that for PCs, which saw efficiency double every 1.52 years from 1975 to 2009 (see Figure 4). The data show significant increases in computational efficiency even during the vacuum tube and discrete-transistor eras. From 1946 (ENIAC) to 1958 (when the last of the primarily tube-based computers in our sample came on line), computations per kilowatt-hour doubled every 1.35 years. Computations per kilowatt-hour increased even more rapidly during the shift from tubes to transistors, but the pace of change slowed during the era of discrete transistors.

In the recent years for which we have more than a few data points (2001, 2004, 2008, and 2009), there is a factor of two or three separating the lowest and highest estimates of computations per kilowatt-hour, which indicates substantial variation in the data in any given year. This variation is partly the result of including different types of computers in the sample (desktops, servers, laptops, and supercomputers), but the differences tend to be swamped by the rapid increase in performance per computer over time, which drives the results.

Explaining These Trends

Even current computing technology is far from the minimum theoretically possible energy used per computation. In 1985, the physicist Richard Feynman analyzed the electricity needed for computers that use electrons for switching and estimated that there was a factor of $10^{11}$ improvement that was theoretically possible compared to computer technology at that time. Since then, performance per kilowatt-hour for computer systems has improved by a factor of $4 \times 10^4$ based on our regressions, but there is still a long way to go with current technology before reaching the theoretical limits—and that doesn’t even consider the possibility of new methods of computation such as optical or quantum computing.

For vacuum-tube computers, both computational speed and reliability issues encouraged
computer designers to reduce power use. Heat reduces reliability, which was a major issue for tube-based computers. In addition, increasing computation speeds went hand in hand with technological changes (such as reduced capacitive loading, lower currents, and smaller tubes) that also reduced power use. And the economics of operating a tube-based computer led to pressure to reduce power use, although this was probably a secondary issue in the early days of electronic computing.

For transistorized and microprocessor based computers, the driving factor for power reductions was (and is) the push to reduce the physical dimensions of transistors, which reduces the cost per transistor. To accomplish this goal, power used per transistor also must be reduced; otherwise the power densities on the silicon rapidly become unmanageable.
Per-transistor power use is directly proportional to the length of the transistor between the source and drain, the ratio of the transistor length to the electrons’ mean free path between collisions, and the total number of electrons in the operating transistor, as Feynman pointed out. Shrinking transistor size therefore resulted in improved speed, reduced cost, and reduced power use per transistor. Power use is driven by more than just transistor size, however. Computer systems include losses in power supplies and electricity used by disk drives, network cards, and other components. And the energy efficiency associated with these components does not necessarily improve at rates comparable to the trends identified in this article. More research is needed to understand the relative contributions of these different components to progress in the electrical efficiency of computer systems as a whole.

**Historical and Future Implications**

These trends have been critical for the historical development of mobile computing. They also have implications for the total power used by computers over time and for the availability and ubiquity of battery powered mobile computing devices in the future.

**Historical Development of Mobile Computing**

The trends identified in this research have strongly affected the development of mobile computing technologies because these devices are constrained by battery storage. As computations per kilowatt-hour increase (holding battery capacity constant), more mobile devices became feasible. Performance and efficiency improvements are inextricably linked, and in some sense, mobile computing is the inevitable result of long-term improvements in computations per kilowatt-hour.

The most visible beneficiaries of these trends have been laptop computers, cellular phones, and personal digital assistants. For example, sales of laptop computers (which use significantly less power than desktop machines) exceeded sales of desktops for the first time in 2009, according to IDC data, demonstrating that portable computers are displacing desktop machines in many applications. This development would not have been possible without long-term improvements in computational efficiency because battery technologies have not improved in the past nearly as rapidly as semiconductor technologies.

**Total Electricity Used by Computing Equipment**

The total electricity used by computers is not just a function of computational efficiency; the total number of computers and the way they are operated also matter. Table 1 shows the total number of PCs in 1980 and 1985, estimated from historical shipments (see http://arstechnica.com/old/content/2005/12/total-share.ars), and for 1996, 2000, and 2008 as estimated by IDC.

The table shows that the installed base of personal computers doubled on average about every three years between 1980 and 2008. Performance growth per computer has just about cancelled out improvements in performance per kilowatt-hour in the PC era (the doubling times are both about 1.5 years), so we would expect total PC electricity use to scale with the number of PCs. However, that simple assessment does not reflect how the technology has evolved in recent years.

First, the metric we analyzed here focuses only on the peak power use and performance of computers—it says nothing about the energy use of computers in other modes (which are the dominant modes of operation for most servers, desktops, and laptops). Servers in typical business applications approach 100% computational load on average for...
only 5 to 15% of the time,\textsuperscript{17} and desktop and laptop machines usually have even lower utilization.

Second, laptop computers (which typically use one-third to one-fifth of the power of a comparable desktop, as shown in Table S2 in the Web Extra appendix) have started to displace desktops in many applications. That trend is confirmed by the data in Table 1. Liquid crystal display (LCD) screens, which use about one-third of the power of comparable cathode-ray tube (CRT) monitors, have largely displaced CRTs for desktop computers since 2000. More recently, LCD screens have seen significant efficiency improvements with the advent of LED backlighting.

Finally, the EPA’s Energy Star program for office equipment has had a substantial impact on the electricity used by this equipment since its inception in the early 1990s,\textsuperscript{18,19} particularly when computers are idle (which is most of the time). The program has promoted the use of low-power innovations in desktop machines that were originally developed for laptops.

A complete analysis of electricity used by computing over time would tally installed base estimates for all types of computers and correlate those numbers with measured power use and operating characteristics for each computer type over all their operating modes, including the low-power modes promoted by Energy Star.

**Implications for the Future**

The computer industry has been able to sustain rapid improvements in computations

---

**Table 1. Installed base estimates for desktop and laptop computers (millions of units).**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Desktop PC</td>
<td>US</td>
<td>80.4</td>
<td>151.3</td>
<td>194.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Western Europe</td>
<td>58.4</td>
<td>92.3</td>
<td>130.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Japan</td>
<td>12.4</td>
<td>21.4</td>
<td>30.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asia Pacific (excluding Japan)</td>
<td>34.0</td>
<td>71.1</td>
<td>249.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Latin America</td>
<td>10.5</td>
<td>26.6</td>
<td>79.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Canada</td>
<td>8.5</td>
<td>16.0</td>
<td>20.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Central and Eastern Europe</td>
<td>7.1</td>
<td>13.3</td>
<td>47.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middle East and Africa</td>
<td>4.2</td>
<td>9.6</td>
<td>30.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>215.5</td>
<td>401.7</td>
<td>784.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portable PC</td>
<td>US</td>
<td>14.3</td>
<td>30.9</td>
<td>121.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Western Europe</td>
<td>6.4</td>
<td>14.9</td>
<td>103.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Japan</td>
<td>5.9</td>
<td>17.2</td>
<td>38.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asia Pacific (excluding Japan)</td>
<td>2.8</td>
<td>7.3</td>
<td>78.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Latin America</td>
<td>0.6</td>
<td>1.5</td>
<td>18.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Canada</td>
<td>0.8</td>
<td>2.8</td>
<td>12.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Central and Eastern Europe</td>
<td>0.4</td>
<td>0.8</td>
<td>25.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middle East and Africa</td>
<td>0.4</td>
<td>1.1</td>
<td>15.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>31.6</td>
<td>76.5</td>
<td>414.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Grand total**

<table>
<thead>
<tr>
<th></th>
<th>2.1</th>
<th>23.1</th>
<th>247.1</th>
<th>478.2</th>
<th>1198.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index (1980 = 1)</td>
<td>1.00</td>
<td>1.1</td>
<td>1.17</td>
<td>2.27</td>
<td>5.68</td>
</tr>
<tr>
<td>Average annual growth since 1980</td>
<td>61%</td>
<td>35%</td>
<td>31%</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>Doubling time since 1980 (years)</td>
<td>1.45</td>
<td>2.33</td>
<td>2.56</td>
<td>3.06</td>
<td></td>
</tr>
<tr>
<td>Index (1985 = 1)</td>
<td>1.00</td>
<td>1.07</td>
<td>2.07</td>
<td>52.0</td>
<td></td>
</tr>
<tr>
<td>Average annual growth since 1985</td>
<td>24%</td>
<td>22%</td>
<td>19%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doubling time since 1985 (years)</td>
<td>3.21</td>
<td>3.43</td>
<td>4.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The data for 1996 to 2008 originated from David Daoud at IDC.\textsuperscript{16} The installed base in 1980 and 1985 is based on historical shipments data from http://arstechnica.com/old/content/2005/12/total-share.ars and an assumed CPU lifetime of five years, which is comparable to IDC’s assumptions.
per kilowatt-hour over the past 65 years, and we fully expect those improvements to continue in coming years. This research suggests that doubling of computations per kilowatt-hour about every year and a half is the long-term industry trend. Because of the large remaining potential for efficiency, we believe that achieving faster rates of improvement is within our grasp if we make efficiency a priority and focus our efforts on a holistic compute system approach, constantly revisiting the notion of what Amory Lovins of Rocky Mountain Institute calls “clean slate, whole system redesign.”

Whether performance per CPU can grow for many years more at the historical pace is an ongoing debate in the computer industry,20 but near-term improvements, such as 3D transistors, are already “in the pipeline.” At this juncture, continuing the historical trends in performance (or surpassing them) is dependent on significant new innovation comparable in scale to the shift from single core to multi-core computing. Such innovation will also require substantial changes in software design,21 which is a relatively new development for the IT industry and is another reason why whole-system redesign is so critical to success.

The most important future effect of these trends is that the power needed to perform a task requiring a fixed number of computations will fall by half every 1.5 years, enabling mobile devices performing such tasks to become smaller and less power consuming and making many more mobile computing applications feasible. Alternatively, the performance of mobile devices could continue to double every 1.5 years while maintaining the same battery life (assuming battery capacity doesn’t improve).

These two scenarios define the range of possibilities. Some applications (such as laptop computers) will likely tend toward the latter scenario, while others (such as mobile sensors and controls) will take advantage of increased efficiency to become less power hungry and more ubiquitous.

Conclusions
The performance of electronic computers has shown remarkably steady growth over the past 65 years, a finding that is not surprising to anyone with even a passing familiarity with computing technology. In the personal computer era, performance per computer has doubled approximately every 1.5 years, a rate that corresponds with the popular interpretation of Moore’s law. What most observers do not know, however, is that the electrical efficiency of computing (the number of computations that can be completed per kilowatt-hour of electricity) also doubled about every 1.5 years over that period.

Remarkably, the average rate of improvement in the electrical efficiency of computing from ENIAC through 2009 (doubling approximately every 1.6 years) is comparable to improvements in the PC era alone. This counterintuitive finding results from significant increases in power efficiency during the tube computing era and the transition period from tubes to transistors, with somewhat slower growth during the discrete-transistor era.

The main trend driving increased performance and reduced costs in recent decades, namely smaller transistor size, also tends to reduce electricity use, which explains why the industry has been able to improve computational performance and electrical efficiency at similar rates. Similarly, reduced capacitive loading, lower currents, and smaller tubes helped vacuum-tube computers significantly improve their energy efficiency over time. The existence of laptop computers, cellular phones, and personal digital assistants was enabled by these trends, which if they continue, presage continuing rapid reductions in the power consumed by mobile computing devices, accompanied by new and varied applications for mobile computing, sensors, and controls.

Acknowledgments
This report was produced with grants from Microsoft and Intel and independent review comments from experts throughout the industry. All errors and omissions are the responsibility of the authors alone. We thank Rob Bernard of Microsoft and Lorie Wigle of Intel for their financial support of this project. We also give special thanks to Ed Thelen of the Computer History Society, who endured innumerable questions about early computers and was remarkably patient in sharing his knowledge and insights (he also put us in touch with Bill Newman, Robert B. Garner, and Rick Dill, who educated us about the power use of early transistors compared to tubes). We would like to thank Kaoru Kawamoto for supplying recent measured data for PC active power; John Goodhue at SiCortex for supplying and explaining data on his company’s largest supercomputer;
Wu-chun Feng at Virginia Tech for supplying data on the Green Destiny supercomputer; Mark Monroe of Sun Microsystems for supplying power data on three older Sun servers; Gordon Bell of Microsoft for supplying power data on some early DEC machines; Saul Griffith and Jim McBride for useful discussions about information theory and power use; Jay Young of Cray for measured power use of the Cray CS6400; Erik Klein (collector of historical computers) for allowing Koomey to measure the power use of his Altair 8800, Apple II, and Osborne I machines; Mary Ann Piette, Richard E. Brown, and Bruce Nordman at Lawrence Berkeley National Laboratory (LBNL) for digging through their archives for relevant materials; and Erik Brynjolfsson at MIT for providing helpful guidance about the regressions. Other colleagues at LBNL allowed us to meter their computers, and for that access we are also grateful. We would also like to thank Bill Nordhaus of Yale for his superbly documented historical analysis of computing performance and for his help in understanding its subtleties. Finally, we would like to thank David Daoud and Tom Mainelli of IDC for generously sharing their data on sales and installed base for PCs and sales of monitors, and Vernon Turner of IDC for facilitating that data sharing.

References and Notes

10. All doubling times in the text are derived from the regression analyses described and documented in the Web Extra appendix and Table S1. See http://doi.ieeecomputersociety.org/10.1109/MAHC.2010.28.  

Jonathan Koomey is a consulting professor at Stanford University. He is one of the leading international experts on the economics of reducing greenhouse gas emissions and the effects of information technology on resource use. He is also the author of *Turning Numbers into Knowledge: Mastering the Art of Problem Solving*, 2nd edition (Analytics Press, 2008). Koomey has a PhD from the Energy and Resources Group at the University of California at Berkeley. Contact him at jgkoomey@stanford.edu.

Stephen Berard is a senior program manager in the Windows kernel group at Microsoft, where he was responsible for energy efficiency and power management features in Windows. His primary focus is on reducing energy needs and operating costs through innovations in Windows power management. Berard has a BS in computer science from the University of Rhode Island. Contact him at sberard@microsoft.com.

Marla Sanchez is a PhD candidate in engineering and public policy at Carnegie Mellon University. Her research areas include assessing potential for environmental life-cycle reductions in products, assessing environmental impacts and policy strategies for the information sector, and advancing standards and procedures to support use-phase efficiency and life-cycle carbon reporting efforts. Sanchez has a MS in energy and resources from the University of California at Berkeley. Contact her at mcsanche@cmu.edu.

Henry Wong is a senior staff platform technologist at Intel. He has more than 16 years of industry experience in digital and mixed signal processor development and more than five years of experience in technology development and enablement of high-efficiency and high-reliability power-conversion techniques for Intel’s Itanium and Xeon processor platforms. Wong has a BS in semiconductor physics and econometric modeling from Yale University. Contact him at henry.j.wong@intel.com.