Flattening the World Efficiently: Digital Sustainability for the 21st Century

Benjamin C. Lee
School of Engineering and Applied Sciences
Harvard University

Word Count: 2,072 (excl. bibliography)

Industrial Ecology in the Digital Era

Information technology (IT) infrastructure will drive globalization and growth in the world economy of the twenty-first century. These technologies expose comparative advantages and increase productivity, effectively flattening the competitive landscape (Friedman, 2006). While the economic benefits of digital adoption are often cited, the environmental impact of these technologies remains unclear. Policy research often emphasizes the potential of IT to "strengthen economic development with less pressure on the environment" (Alakeson, 2003). Examples often include potential emission reductions from substitution effects between digital and traditional practices. Despite their oft cited environmental benefits, IT infrastructure consumes significant amounts of energy and the corresponding environmental costs must also be considered. Energy efficiency is particularly critical since the bulk of electricity generation requires fossil fuels. As the IT industry matures from advances in and adoption of digital practices, the international community must ensure these processes are energy efficient by establishing a framework for monitoring and communicating their net environmental impacts.

Information technology offers unprecedented opportunities in environmental knowledge management (EKM) systems to "connect data, analysis, and people...[and] formalize industrial ecology in a business setting." (Wernick, 2003) This paper will argue EKM should be applied to the IT industry itself with particular emphasis on data centers and other centralized computational resources. Monitoring and management frameworks should be implemented by both providers and users of IT infrastructure to assess and communicate the environmental impact of their products and digital practices,

respectively. This paper identifies significant open questions within industrial ecology for the digital era and proposes guidelines for collecting the data to answer them.

Centralization of Computing Resources

High performance computing resources are increasingly centralized in data centers, web servers, and other computational clusters. Individual electronic devices often serve as access points to these centralized resources and, as a result, are no longer computational workhorses. In future, digital consumers will access tools on the Web, use them on the Web, and store business data on the Web (Friedman, 2006). As application developers such as Microsoft, Google, and Salesforce.com increase their web-based offerings, computational requirements will shift from personal computers toward centralized servers operated by these firms. This trend is further reinforced by the prevalence of hand-held devices able to access and utilize these web-based services.

These trends in centralization, driven by the commoditization of compute servers and economies of scale, suggest environmental and energy effects from IT infrastructure are most easily observed in aggregate from data centers. Market analysts project 14 to 15 percent server market compound annual growth with much of this growth attributed to mega data centers with more than 25 thousand servers (Turner, 2006). Furthermore, these analysts project the continued commoditization of server hardware with more servers sold at lower prices. Such trends will further contribute to the scale out of computational resources within the larger data centers. Mega data centers will likely represent an increasingly large fraction of energy consumption by IT infrastructure, requiring power for both compute hardware and HVAC[1] cooling infrastructure. Cooling power is often estimated at 50 to 75 percent of server power (Turner, 2006).

This shift in energy consumption toward centralized resources is complemented by increasingly advanced power management techniques for individual electronic devices. For example, voltage and frequency scaling lowers energy consumption by reducing

[1] HVAC: heating, ventilating, air conditioning

microprocessor performance during periods of low utilization. Such periods will become increasingly common as computation shifts toward centralized resources and microprocessors spend more time waiting for responses from the network. These waiting periods expose greater opportunities for aggressive power reduction techniques, thus reducing the environmental impact of energy consumption from individual devices. The net effect of these trends is a shift in energy consumption from individual electronic devices to the data centers providing these devices with web-based applications.

Impact Assessments from IT Providers

The providers of IT infrastructure are best able to provide energy efficiency data for their products. A typical data center is comprised of equipment to provide compute and storage capabilities in a scalable manner as well as support infrastructure for network connectivity, reliability, and cooling. Data center efficiency must be assessed at various levels, including the microprocessor (e.g., servers), system (e.g., networking and storage), and facility (e.g., HVAC) levels. Energy consumption is increasingly significant with compute hardware requiring on the order of megawatts of power. As Moore's Law continues to drive greater hardware densities, we will also observe increasing power density and the corresponding thermal effects.[2] Thus, HVAC energy consumption is also significant and likely to increase. Data center management is very much an area of industrial engineering and operations research, not just an area of electrical engineering and computer science. Therefore, the technical advances in IT must be coupled with the effective application of known techniques in optimization and industrial ecology.

Energy consumption in the digital era is a pressing concern when most of this energy is still generated from fossil fuels such as coal, oil, or natural gas. Less energy consumed by IT infrastructure implies fewer power plants and fewer carbon emissions. Inefficiencies in existing IT infrastructure provide opportunities for energy optimization at all levels in the system. At the server level, advances in low power design and

[2] Moore's Law is the empirical observation that the number of transistors on an integrated circuit doubles every 18 months. This observation implies increasingly small electronic devices.

multiprocessor architectures reduce energy consumption (Williams, 2006). Multiprocessor architectures achieve higher performance for a given energy budget while more successfully controlling the thermal effects of system utilization. At the system level, research in workload scheduling examines techniques for assigning applications to the servers able to most efficiently perform the computation (Chase, 2001). At the facility level, best practices are being developed to counter inefficiencies in power conversion and cooling (Sullivan, 2006). Cool air is usually circulated through the system to remove server generated heat, but inefficient air flows often result in over-provisioning of HVAC resources.

A unified framework is needed to assess these optimizations in technical design and data center operation. Such a framework should specify metrics for comparing design efficiencies, methodologies for obtaining these metrics, and application workloads for which these metrics are obtained. Efficiency metrics for the data center should include measures of performance, energy consumption, and utilization (Papadopoulos, 2006). For example, efficiency metrics for microprocessor design should examine the rate at which energy is consumed to provide a particular level of performance. This metric should be scaled by utilization to reflect efficiency under typical usage patterns instead of theoretical peak capacity. To facilitate comparison, these metrics must be obtained using standardized methodologies. For example, the rate of energy consumption could come from a variety of sources: (1) manufacturer reported peak, (2) user measured peak, and (3) user measured average. While user measured averages are most useful for assessing energy efficiency, they are often most difficult to measure and must be obtained from systems executing standardized workloads. These workloads must be representative of modern software and stress various components of the system. Candidates for these workloads may arise from enterprise applications, search algorithms, and data mining heuristics. The comprehensive framework that integrates these ideas will likely follow precedence set by, for example, the United States Environmental Protection Agency for energy efficient electronic appliances (Energy Star, 2006) or the German Federal Environmental Agency for environmentally friendly products (Blue Angel, 2006).

Impact Assessments from IT Users

To assess fully the environmental impact of IT infrastructure, we must also consider trends in technology adoption and their impact on traditional business practices. Productivity and efficiency gains are undisputed benefits from IT adoption, but claims of a net environmental benefit from the adoption of digital business processes are often more difficult to validate. In particular, substitution effects arising from IT adoption are often too easily assumed to mitigate the environmental impact of business operations. Predictions of paper-less offices were very common during the rise of digital media, but "electronic media are not so much a substitute for as a supplement to printed or other media, thus tending to increase environmental impacts. The risk exists for incomplete substitution and for the additional use of electronic media next to conventional media." (Fichter, 2003) More recently popular is the idea of telecommunications and network connectivity as substitutes for traditional business travel. Again illustrating incomplete substitution, these technological advances complement traditional practices by producing "faster growth in telecommunications than in travel, resulting in an increasing share of [business interactions attributed] to telecommunications, but with continued growth in travel in absolute terms." (Mokhtarian, 2003) Thus, incomplete substitution and complementary practices obscure the net environmental effects of IT adoption.

Users may shed light on these effects by communicating the degree to which IT has been incorporated into its operations. This supplementary monitoring of digital adoption may then be correlated against data within a user's existing environmental assessments. For example, the frequency of video conferencing may be correlated against that of business travel where the latter is likely already reported in the context of carbon emissions accounting. When this data is drawn from multiple companies and industries in different stages of digital adoption, trends in substitution effects may be more clearly assessed. The international community should construct a framework to formalize areas of technology adoption and the corresponding areas of potentially displaced traditional practices. For example, technologies could include digital media and telecommunications while their traditional analogues could include paper media and

travel. Within this framework, the degree of digital substitution should be communicated to the public with special recognition granted to users most effectively leveraging IT infrastructure to reduce or slow the growth of less environmentally friendly operations. Precedence for such recognition includes the Financial Times' list of the world's most respected companies, which accounts for environmental reputation and efforts in sustainability when compiling its rankings. (Maitland, 2001)

Users may also indirectly report the effects of IT adoption via associated energy consumption (Fichter, 2003). This approach is attractive as many entities already monitor these numbers through existing frameworks for emissions accounting. Furthermore, energy consumption is a common metric that correlates directly with environmental concerns such as global warming. However, this approach is not guaranteed more effective than supplementary monitoring of digital adoption. Individual users may assess substitution effects from digital adoption, but identifying trends in digital sustainability through energy consumption depends on effective cooperation between providers and users of IT infrastructure. For example, a user might adopt enterprise applications from Salesforce.com. Since these applications run on servers within Saleseforce.com data centers, the web-based application provider must communicate to the user its share of energy consumption for server utilization and cooling. In the absence of such cooperation, significant energy consumption and environmental effects may simply move beyond the boundaries of the monitored user. Due to the demands of such cooperation, frameworks for individual users to communicate IT adoption and substitution trends may be more effective.

Beyond Ecology and Economics

Energy efficient IT infrastructure will limit carbon emissions from electricity generation while benefits of lower utility costs are passed, in part, from providers to users. However, digital sustainability also has significant strategic and socioeconomic implications. For example, supercomputing facilities are effectively high performance data centers. Energy efficient supercomputing drives computational scalability, the basis

for advances in molecular and fluid dynamics simulation. These simulations enable weapon stockpile management without explicit testing and cost effective air/naval defense prototyping. Simulations also reduce the environmental impact of these strategic operations by reducing high impact explosive experimentation and stress testing. Socioeconomic opportunity costs of inefficient technologies should also be considered. For example, India has very successfully developed its information services industry, but inefficient data centers may lead to over-provisioning of utilities and other infrastructure projects in regions around technology hubs (Economist, 2005). The urgent need to strike a balance between urban and rural infrastructure development is illustrated by statistics indicating half of all Indian homes have no electricity (Economist, 2006).

Information technology has a demonstrated ability to connect data, analysis and people, thereby driving global economic growth. However, these benefits are not without environmental costs and the international community must not neglect the broad implications of digital sustainability simply because IT infrastructure's environmental impact is low relative to traditional analogues. On the contrary, global leaders should seize the opportunity to influence the development of a young industry by advocating the monitoring and communication of energy efficiency measured in absolute terms. As efficiency metrics are communicated voluntarily and publicly, IT providers and users will increasingly compete and differentiate based on their efforts in digital sustainability. Environmentally motivated differentiation produces self-reinforcing effects in sustainability that will not only drive unimaginable advances in technology, but also continue to flatten the world and deliver unparalleled economic prosperity.

Bibliography

Alakeson, Vidhya and Wilsdon, James (2003): "Digital Sustainability in Europe." Journal of Industrial Ecology 6(2).

Blue Angel (2006): http://www.blauer-engel.de

Chase, J., Anderson D, et. al. (2001): "Managing Energy and Server Resources in Hosting Centers." Proceedings of the 18th ACM Symposium on Operating System Principles.

Economist (2005): “The Bangalore paradox.” The Economist: 21 April

Economist (2006): “Light and shade.” The Economics: 10 August

Energy Star (2006): http://www.energystar.gov

Fichter, Klaus (2003): “E-Commerce: Sorting Out the Environmental Consequences.” Journal of Industrial Ecology 6(2).

Friedman, Thomas (2006): The World is Flat: A Brief History of the Twenty-first Century. New York: Farrar, Straus and Giroux.

Maitland, Alison (2001): “Due recognition given for effort.” Financial Times: 13 December

Mokhtarian, Patricia (2003): “Telecommunications and Travel: The Case for Complementarity.” Journal of Industrial Ecology 6(2).

Papadopoulos, Greg (2006): “Impacts and Importance of Energy Efficiency: Industry Viewpoint.” CTO, Sun Microsystems: United States Environmental Protection Agency Conference on Enterprise Servers and Data Centers.
http://www.sun.com/aboutsun/environment/epa.jsp

Sullivan, Bob (2006): “Understanding the Impact and Savings Potential for More Efficient Enterprise Servers and Data Centers.” Sr. Consultant, The Uptime Institute: United States Environmental Protection Agency Conference on Enterprise Servers and Data Centers. http://www.sun.com/aboutsun/environment/epa.jsp

Turner, Vernon (2006): “Defining the Landscape: Trends and Forecasts.” Group VP and General Manager of Enterprise Computing, IDC: United States Environmental Protection Agency Conference on Enterprise Servers and Data Centers.
http://www.sun.com/aboutsun/environment/epa.jsp

Wernick, Iddo (2003): “Environmental Knowledge Management.” Journal of Industrial Ecology 6(2).

Williams, Ben (2006): “Maximizing Data Center Efficiencies Through Microprocessor Solutions.” VP of Commercial Business Solutions, AMD: United States Environmental Protection Agency Conference on Enterprise Servers and Data Centers.
http://www.sun.com/aboutsun/environment/epa.jsp