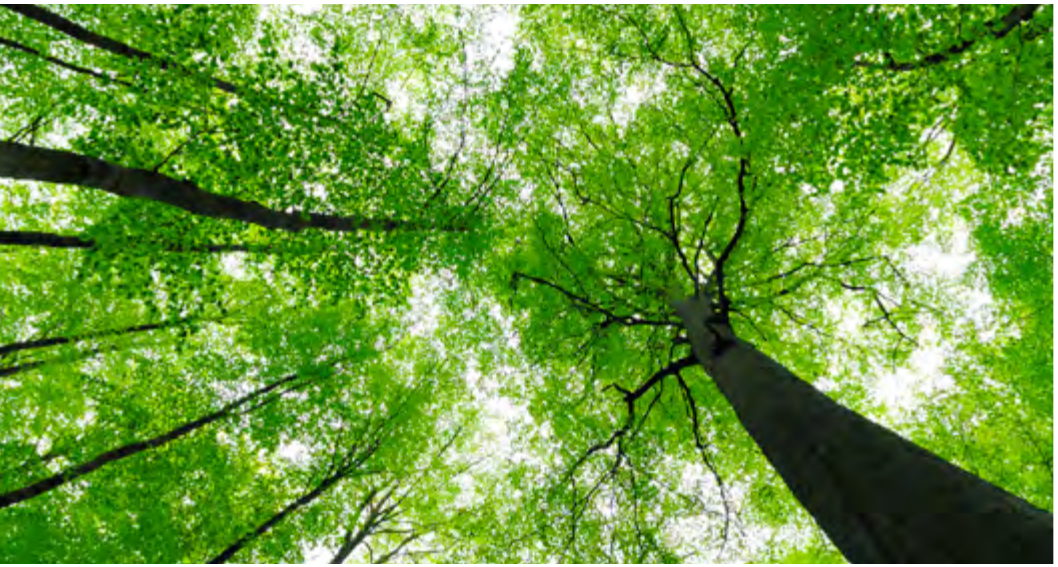
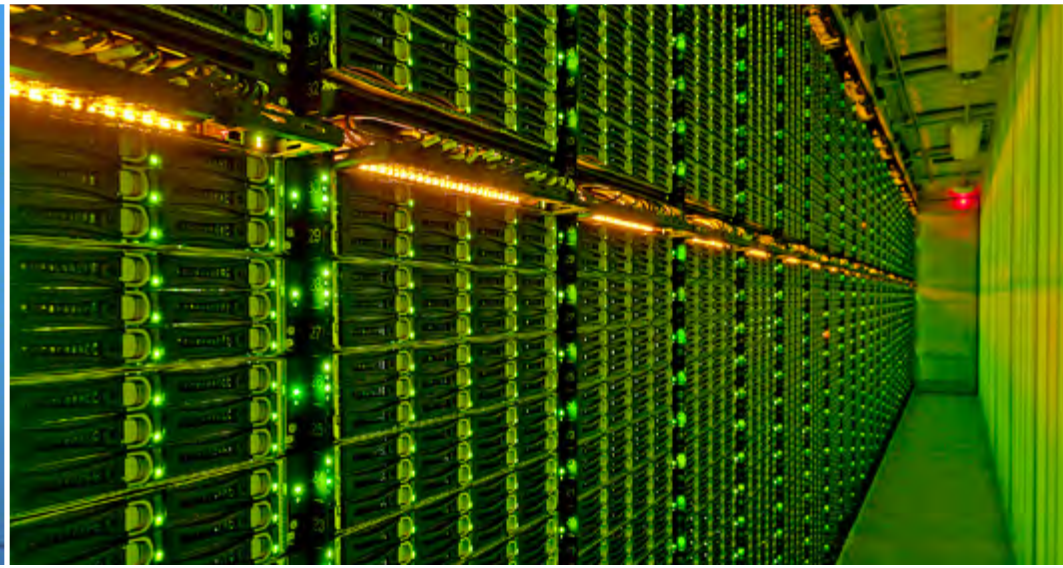
The IT Energy Efficiency Imperative

A white paper for decision makers on the urgency and benefits of embracing IT energy efficiency principles and practices





Microsoft

Contents

Next

The IT Energy Efficiency Imperative

Mark Aggar

Senior Director of Technology Strategy,
Environmental Sustainability, Microsoft

Microsoft Corporation | June 2011

www.microsoft.com/environment

Contents

Executive Summary

Tight budgets, rising energy costs, and limits on electric power availability are hindering the ability of information technology (IT) departments to meet the growing demand for IT services. To mitigate these challenges, it is important for organizations to embrace IT energy efficiency principles and practices to remain productive and competitive in this increasingly digital age.

IT suppliers and vendors are responding to this challenge by producing more energy-efficient computer hardware, creating software that helps improve the energy efficiency of IT operations, and offering improved operational efficiency through public and private cloud computing platforms and services.

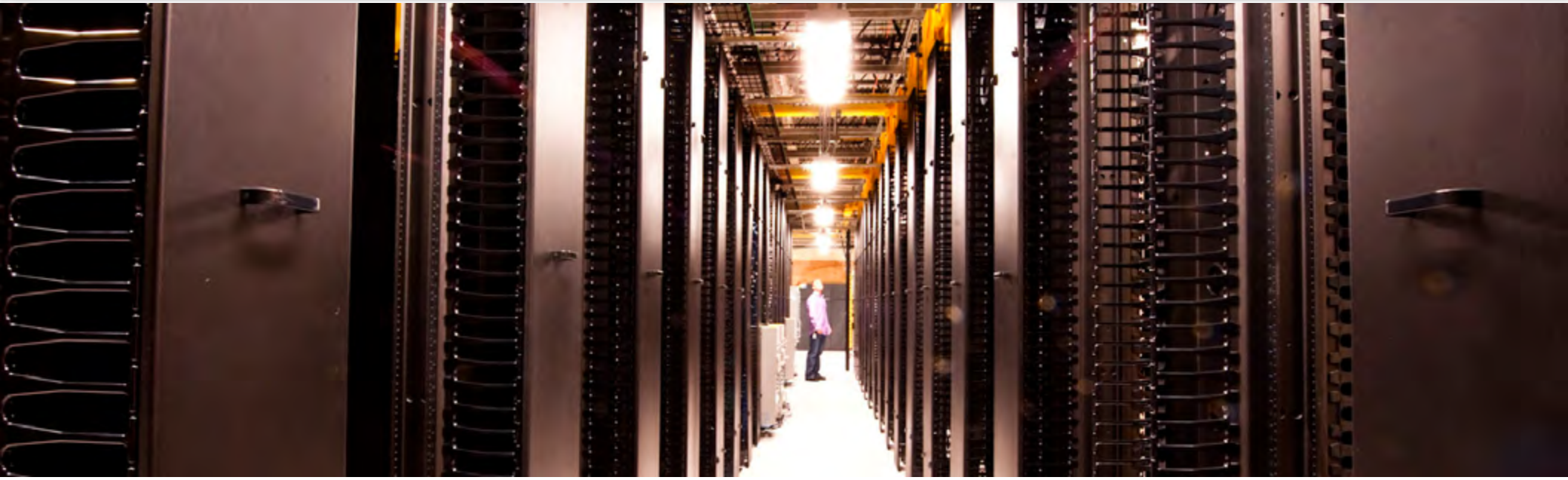
But many IT departments are failing to capitalize on these advances. Lacking the necessary incentives or capabilities to improve IT energy efficiency, they are continuing the decades-old practice of overbuilding computer systems and thereby missing the opportunity to make their IT operations more sustainable and more responsive to organizational needs.

Despite the wide availability of server virtualization and centralized PC power management solutions, only 25% of IT departments have a plan for optimizing IT resource use, increasing energy efficiency, and minimizing the waste generated by their IT operations. As a result, average server utilization remains at historically low levels, and in many organizations, desktop PCs waste as much as 75% of the electricity they consume.

Despite advances in hardware and software, average server utilization remains at historically low levels, and in many organizations, desktop PCs waste as much as 75% of the electricity they consume.

By embracing energy efficiency in areas such as system architecture, hardware provisioning, software design, and operations, organizations can reduce their IT budgets and respond more rapidly to demands for additional services, thereby improving their overall productivity and competitiveness.

As a leading provider of software platforms and solutions, Microsoft is committed to helping customers improve the energy efficiency of their IT systems. We also recognize that manufacturing, operating, and disposing of IT equipment has a significant and growing impact on the environment. Building on capabilities in our core platform offerings—Windows Server, Windows Azure, and Microsoft System Center—that enable energy-efficient IT operations, this paper outlines key strategies, principles, and practices that organizations can adopt to make their IT infrastructure more energy efficient, cost effective, and responsive to business demands.



Introduction

Information technology (IT) is deeply woven into the fabric of modern society, from the way we interact with friends and family to how organizations communicate and collaborate. As the demand for IT—including online services and cloud computing—continues to grow, it is important for decision makers to ensure that their organizations are well positioned to optimize existing IT services and expand them as needed to meet demand.

This is particularly important as fiscal constraints, coupled with the rising cost of energy needed to power IT systems, is making it more difficult for organizations to add new IT capacity. Limited electric power supply, on the grid and within data centers, is also constraining the amount of IT equipment that many organizations can add to their data centers.

The good news is that computer hardware—along with data center power and cooling infrastructure—is becoming more energy efficient and continuing to offer gains in performance and capacity. Software solutions that improve energy efficiency, including server virtualization and centralized power management, are more widely available. Cloud computing infrastructures, both public and private, are offering significant energy efficiency gains compared to traditional IT infrastructures.

Even with these advances, many IT departments struggle with growing IT power demands and high energy costs. A key reason, and a

primary focus of this paper, is poor IT resource utilization. Despite the widespread implementation of server virtualization technologies, about two-thirds of organizations report that less than half of their production environment has been virtualized.[1] Average server utilization is still at or below 15% to 20%[2]—probably no better than it was a decade ago. Many organizations, including the U.S. government, report average server utilization rates of less than half that.[3,4] This is a huge waste of resources, especially considering that 15% of servers in large organizations are completely idle.[5]

Why are IT departments still failing to significantly improve IT resource utilization? Although financial constraints have certainly played a role in slowing down virtualization efforts, particularly in the last couple of years, two related factors appear to be at the root of this phenomenon:

- IT departments do not control all of their organization's IT assets. Most IT departments do not control many of the "mission-critical" applications within their organization. This lack of control results in underutilized hardware that is often ring-fenced and unavailable for use by other applications, thus severely limiting the potential for improving overall IT resource utilization.
- Traditional application designs make it difficult to optimize IT resource allocation. Most applications provide IT operators with little insight into their actual IT resource needs and requirements, which can vary significantly from moment to moment. The applications often rely on dedicated hardware for high availability, and when IT resources are constrained they simply slow down, sometimes with unexpected or unwelcome results.

{ It is important for decision makers to provide their IT departments with the necessary incentives and capabilities to improve the energy efficiency of their operations.

As server performance and storage capacity increase, IT resource utilization and overall IT energy efficiency will continue to decline unless these factors are addressed. The underutilization of IT assets wastes money and takes a toll on the environment that extends far beyond the energy consumed by hardware. Computer manufacturing is a resource-intensive process, many of the materials and energy resources are damaging to the environment when extracted or processed, and some are in increasingly short supply.

Similar issues exist in the client computing environment. For example, many organizations have not implemented centralized PC power management which could reduce their PC energy use by up to 75%.[6] Even when PC power is centrally managed, applications often inadvertently interfere with PC power management capabilities, preventing the machines from saving energy when they are not being used. Like servers in many organizations, PCs also have very low utilization rates.

Improved IT energy efficiency and resource utilization is achievable if hardware and software vendors, consultants, software developers, and IT departments make it a priority. However, decision makers must provide their IT departments with the necessary incentives and capabilities to improve the energy efficiency of their IT operations.

IT Energy Productivity

Ensuring the reliability of mission-critical systems has been a top priority since the early days of IT. As a result, computer systems have routinely been overbuilt to reduce the likelihood of unplanned disruptions due to hardware or software failures or system slowdowns caused by unexpected user demand.

This pattern of overbuilding is a major cause of poor IT energy productivity. For every 100 watts of power consumed in a typical data center, fewer than 3 watts are associated with actual computing.[7] *In other words, a typical data center consumes more than 30 times the energy that it uses to perform computations.* Most of the remaining energy is wasted due to server underutilization. Power and cooling inefficiencies account for the rest.[8]

The potential energy efficiency gains from increased server utilization compared to traditional efficiency measures are illustrated in Figure 1. (Power Usage Effectiveness[9] [PUE] refers to the ratio of the total amount of power used by the data center facility to the power delivered to the IT equipment; Power Supply Efficiency [PSE] refers to the efficiency of conversion from AC to DC power.)

Increasing server utilization clearly offers greater potential to improve a data center's overall IT energy productivity compared to PUE and PSE improvements—far greater than conventional wisdom has recognized. In fact, until average server utilization approaches 50%,[10] improving server utilization offers the biggest gains in IT energy productivity because servers consume a significant amount of energy when idle—typically between 30% and 60% of the power they consume when fully utilized.

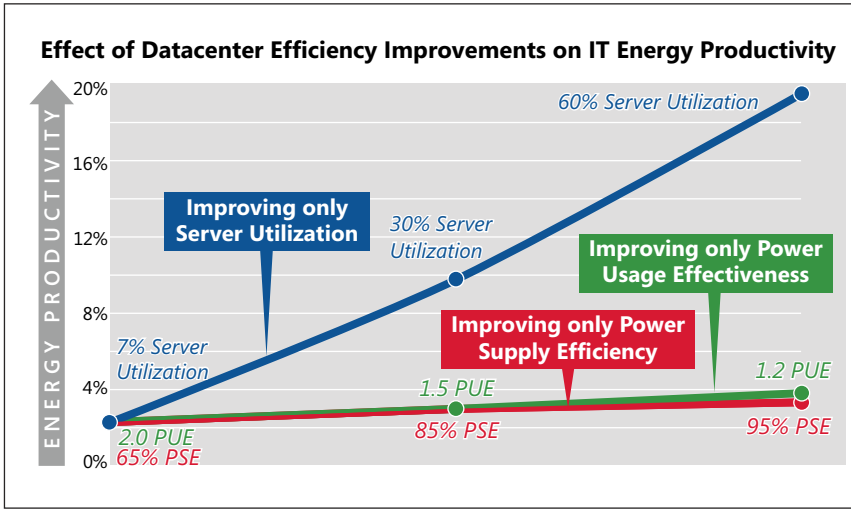


Figure 1. The effects of various data center efficiency measures.

Measuring Utilization

This paper does not delve into how to create an aggregate measure of server utilization that encompasses multiple systems and different subsystems (CPU, memory, disk, and network). This is a complex issue, and the methodology employed depends on the intended use of the utilization data and the sophistication of the management infrastructure and applications. For environments in which "abandoned" servers are common, simply querying whether there is any user activity may be sufficient to determine whether it makes sense to decommission or virtualize a server to improve overall utilization. IT departments with sophisticated operational practices might measure utilization of individual servers and their subsystems in near-real-time to dynamically optimize workload placement.

The Hidden Impacts of Underutilization

The cost of powering chronically underutilized IT equipment can be a significant percentage of an organization's energy bill, and it greatly contributes to data center capacity constraints. These constraints include limits on available utility power, limited power and cooling capacity within the building, and lack of physical space for computers.

The manufacturing of computers that will be underutilized also wastes a significant amount of energy, water, and raw materials (including so called "conflict minerals"). Such equipment typically becomes e-waste within just a few years. In the European Union, for instance, no more than one-third of e-waste is responsibly recycled in a verifiable way.[11] The remaining e-waste often ends up in landfills or is shipped to developing countries, where it is typically dismantled using methods that contaminate the surrounding land, air, and water with toxic metals and chemical compounds, harming the health of unprotected workers and others in the surrounding areas.[12]

Furthermore, given current and projected electric power and resource constraints, failure to significantly improve the utilization of IT equipment will likely limit the ability of IT to help address the world's pressing economic, societal, and environmental challenges.

IT Industry Leadership and Government Oversight

Nearly all computer and data center equipment manufacturers are developing more energy-efficient hardware, and many manufacturers are aiming to meet increasingly stringent Energy Star standards.[13] Computer equipment manufacturers, particularly those that make PCs and monitors, are striving to make manufacturing and disposal processes more environmentally friendly through methods validated by programs such as the EPEAT electronics hardware registry.[14]

Meanwhile, a growing number of software solutions are available to help organizations measure and manage computer energy use and improve hardware utilization through centralized power management, virtualization, and other capabilities. Additionally, many IT vendors and IT departments are working with industry consortiums such as The Green Grid and Climate Savers Computing to develop and promote guidance and tools that help IT energy efficiency efforts.

Cloud computing also has an important role to play in improving IT energy efficiency. Migrating on-premises commodity services to a public cloud computing platform can free up valuable server and power capacity in an organization's data center and reduce the need to invest in IT infrastructure that will be likely be underutilized.

Although the energy efficiency gains of public cloud computing can be realized by organizations of all sizes, a move to the cloud is particularly beneficial for smaller organizations that don't have the user demand to achieve significant utilization of their IT equipment. A recent study by Accenture and WSP Environment & Energy showed that

energy use and carbon emissions are reduced by at least 30% per user when organizations use Microsoft Business Productivity Online Services (such as Microsoft Exchange Online and Microsoft SharePoint Online) or Microsoft Dynamics CRM Online instead of on-premises installations of those applications. For small organizations, the study showed up to a 90% reduction in energy use and emissions.[15] Similar efficiencies are likely achievable when running applications on the Windows Azure cloud computing platform.

Many governments and regulatory bodies are taking action to address the environmental impact of growing IT energy use. For instance, the UK's Carbon Reduction Commitment initiative[16] requires organizations that consume more than 6,000 megawatt-hours (MWh) of electricity per year—the consumption of a data center with about 2,000 small servers[17] —to report their electricity use and purchase carbon allowances. An estimated 20,000 organizations that consume between 3,000 and 6,000 MWh per year will be required to track and report their electricity usage and carbon emissions.

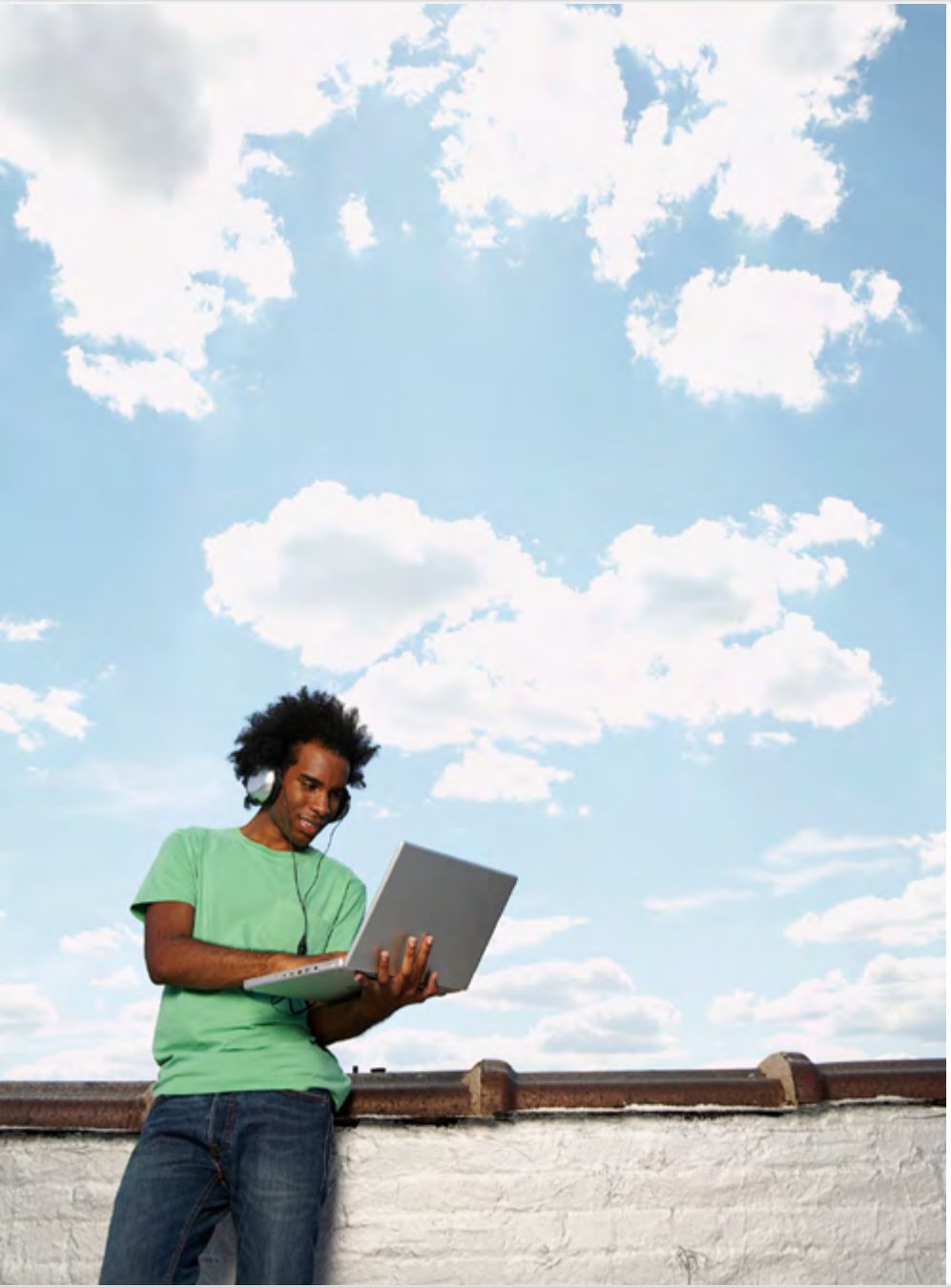
Many governments are also strengthening regulations governing e-waste.[18] In addition to existing laws based on European Union directives (such as WEEE and RoHS[19]) that are aimed at manufacturers of electronic goods, most municipalities now prohibit the disposal of e-waste through standard waste streams and require it to be recycled. The risk of being associated with "e-waste criminals" who dump broken electronics in developing countries is leading many large organizations to require assurances from recyclers that discarded equipment will be recycled and disposed of in an environmentally sound manner.[20]

Power vs. Energy

When discussing IT energy efficiency, it is important to understand the difference between electrical power and electrical energy. Electrical power is the rate at which electricity is consumed (or generated) at any given instant and is measured in watts or multiples of watts—kilowatts, megawatts, and so on. Electrical energy is the quantity of electricity consumed (or generated) over a given time period—minute, hour, year, and so on—and is typically reported and billed in kilowatt-hours (kWh). Thus, a server rack that consumes a constant 20 kilowatts will use 175,200 kWh (175.2 MWh) over the course of a year (8,760 hours).

The distinction is important because data centers are most often constrained by available electrical power infrastructure. The developed world is generally able to produce more energy than it can consume in a year, but there are often shortages of power during peak times.[21] If the demand for power at any given instant exceeds supply—whether due to a lack of generating capacity or power transmission capacity between the generator and the consumer—users will experience a "brownout" that causes computers and other electronic devices stop working.

Even a small reduction in power draw—by, for example, decommissioning unneeded servers—can result in significant energy and cost savings over time. Investing in more efficient equipment with a lower average power draw can also save on construction and equipment costs for electrical and cooling systems and reduce overall energy costs. At scale, these efficiency improvements can reduce the amount of power generation and transmission capacity needed on the electrical grid.



Operating IT as a Utility

Most large IT departments have been trying to improve server utilization for some years through the use of server virtualization technology. But many of these efforts have stalled due to financial constraints, organizational politics, and a shortage of sufficiently skilled IT staff.

Building a consolidated IT infrastructure often requires a significant up-front investment of time and money, and owners of mission-critical applications are sometimes reluctant to migrate applications to shared IT infrastructure due to concerns about reliability. Unfortunately, these applications typically represent a large portion of the IT infrastructure and often have low average utilization rates. Budgeting processes and long lead times between ordering and installing servers often compound the problems by driving business units to overbuy equipment far ahead of actual needs.

In 2010, only one in four IT departments had a plan for further optimizing IT resource use, increasing energy efficiency, and minimizing waste generated by their operations, according to a report by Forrester Research. Improving IT infrastructure energy efficiency was a critical priority for just 12% of companies.[22] Another recent study found that many IT departments did not have sufficient budget to improve the energy efficiency of their IT systems, and that senior management did not view IT energy efficiency as a high priority.[23]

IT departments must be empowered with the appropriate incentives and resources to transition from simply providing "lodging" for other business units' servers to offering "IT as a utility"—through centralized computation-and-storage services hosted in private or public clouds. Operating IT in this manner can dramatically improve equipment

utilization and energy efficiency, as well as significantly reduce IT costs across the entire organization.

Centralized computation-and-storage services are at the heart of so-called public and private clouds. Public clouds consist of IT infrastructure that is shared among many organizations and improves energy efficiency through economies of scale, multi-tenancy, and the motivation of the cloud provider to improve its bottom line without sacrificing reliability or security. In fact, for all but the most sophisticated IT operations, public cloud computing from a reputable vendor, when used appropriately, is almost certainly the most secure, reliable, and cost-effective method of obtaining IT services.

Private clouds—typically installed on premises and operated by and for the exclusive use of a single organization—share many of the same attributes, although they are generally less energy efficient than public clouds because of their smaller scale and because most organizations are less adept at running IT systems than specialized vendors. However, private clouds do have one important advantage over public clouds: migrating data and applications (such as virtual machines) is typically much faster and easier given the physical proximity between the existing IT infrastructure and the private cloud.

In either scenario, consuming IT services from a cloud enables business units to pay for only what they use, which encourages them to be more prudent in their use of IT resources. As the lower costs—compared to a dedicated infrastructure—attract more business units, the cloud will become even more cost effective and energy efficient.

Educating business units on the true cost of running their applications and teaching them how to optimize and reduce IT resource usage are important tasks for an organization that wants to run its IT as a utility. For private clouds, IT departments must work closely with business units to accurately forecast demand and avoid wasteful overbuilding or potentially catastrophic underbuilding of the supporting infrastructure. Of course, any forecasting exercise will be imperfect, but a centrally controlled private cloud allows any excess resources to be used by other applications rather than sitting idle.

IT departments that embrace energy efficiency can stretch their budgets and respond more effectively to demands for additional services, thereby improving their organization's productivity and competitiveness.

IT departments that focus on energy efficiency as a core practice can stretch their budgets and respond more rapidly to demands for additional services, thereby improving their organization's productivity and competitiveness. By embracing IT energy efficiency in areas such as system architecture, hardware provisioning, software design, and operations, organizations will almost certainly realize significant energy and cost savings up and down the IT stack.

Applications: The Missing Link

Most IT energy efficiency efforts have traditionally focused on physical infrastructure—deploying more energy-efficient computer hardware and cooling systems, using operating system power management features, and reducing the number of servers in data centers through hardware virtualization.

But a significant amount of IT energy inefficiency stems from how applications are designed and operated. Most applications are provisioned with far more IT resources (servers, CPU, memory, etc.) than they need, as a buffer to ensure acceptable performance and to protect against hardware failure. Most often, the actual needs of the application are simply never measured. This practice naturally results in poor hardware utilization across the data center.

While virtualization can help improve hardware utilization to a certain degree, many organizations find that utilization is lower now than before they started virtualizing. Applications in virtualized servers are often just as idle as they were when they ran on dedicated infrastructure because their use of IT resources is not scaled dynamically. The result is more instances of active virtual machines than necessary—often referred to as *virtual server sprawl*—and virtual machines being allocated more IT resources than they really need. The problem is compounded by the fact that hardware performance generally grows faster than virtual server consolidation ratios, so with each new generation of hardware, utilization actually decreases.

For the enormous number of existing applications, virtualization management systems can help optimize the placement and configuration of virtual machines to improve utilization of the virtualized server

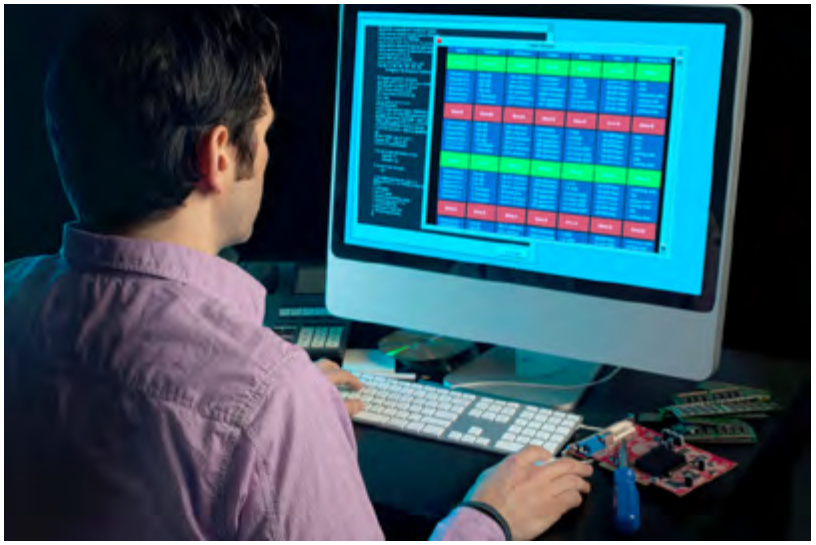
infrastructure. Systems management solutions can also be used to report and even cap the electric power draw of the host infrastructure, effectively constraining an application's use of IT resources.

But fundamental issues remain. Applications are generally not instrumented to expose metrics that could help IT management systems determine the type and quantity of IT resources required to meet certain performance criteria, and they do not provide mechanisms to externally control their use of resources.

Administrators must often guess the quantity and type of IT resources an application needs to operate with acceptable performance. This uncertainty is caused by lack of information about how an application consumes resources, the likely demand for the application

over time, and the manner in which resource use scales with demand.

Developers can significantly reduce such uncertainty by instrumenting their applications to expose usage and performance metrics that can be used to dynamically assign and withdraw IT resources as needed. In recognition of this need, the Marlowe project[24] at Microsoft Research is experimenting with a framework to optimize IT resource allocation for applications by using developer-instrumented usage and performance metrics, in addition to more abstract utilization metrics such as overall CPU utilization.



When these metrics are correlated with the underlying hardware performance, it is possible to dynamically assign the appropriate quantity and type of IT resources to operate the application with desired performance characteristics (as specified by a service level agreement) at any given scale of use. Past usage data can also be analyzed to provide near-real-time use forecasts to assist with IT resource provisioning for the application. Excess capacity can be temporarily repurposed for "opportunistic" computing tasks (such as batch jobs) until it is needed again, or it can be shut down entirely if there is no additional work or if there is a need to reduce energy consumption.

In addition, breaking down applications into fine-grained units of service delivery—as demonstrated by another Microsoft Research initiative, the Orleans project[25]—enables far more effective IT resource usage and control than is typically possible with large, monolithic application components and services. The higher level of abstraction away from the computing resources that Orleans provides allows developers to focus on business value rather than on the complexities of scaling and reliability, and it should markedly increase overall IT resource utilization.

Because the demand for many applications is initially unknown and unpredictable, developers should consider implementing administrator controls that can limit the number of simultaneous users and perhaps dynamically degrade the fidelity of the application experience if cost containment is important or if IT resources are limited. In this way, operating costs can be better managed, and active users of an application will have a predictable experience rather than experiencing random slowdowns resulting from more simultaneous active users than the allocated IT resources can accommodate.

Applications should also be able to defer noncritical work such as batch processing and other maintenance tasks when specified by the IT operator. This will provide additional "virtual capacity" for other applications that have unanticipated spikes in demand. It will also provide additional options for IT departments to temporarily reduce power use, such as when responding to a demand response event from a utility.[26]

Applications designed with energy and resource efficiencies in mind will be more reliable, cost effective, and easier to manage.

Applications that are designed with these IT efficiency goals in mind are also generally much easier to operate in degraded or partially recovered states during disaster recovery situations. Similarly, applications do not need to have dedicated capacity to sustain service during maintenance events. IT resources can be claimed and released as required, keeping the costs of maintenance events to a minimum.

It can also be cost effective for developers and testers of resource-intensive applications to invest in increasing the amount of work produced per watt-hour, with the goal of making energy use scale roughly linearly with useful output. In addition, nearly all applications should be tested to ensure that they do not waste energy while performing little or no useful work.

Clearly, for applications that will be rarely used or are experimental in nature, these investments are probably not necessary. Existing management infrastructures are more than sufficient to take care of scaling, and inadvertently starving these applications of IT resources on a temporary basis will not cause significant harm. However, for the many applications that are likely to be long lived and periodically resource intensive, designing them so their use of IT resources can be dynamically scaled based on demand and other constraints can easily pay off.

Energy Efficiency in Perspective

Although energy efficiency is important, it doesn't pay if it significantly reduces productivity, performance, and reliability. Some important systems are underutilized by design. For example, no one would seriously suggest sharing fire trucks among airports many miles apart to increase their utilization.

Understandably, IT pros are highly cautious and will advise against operational practices that introduce real or perceived risk. However, the vast majority of applications are not going to cause significant harm if they are oversubscribed or unavailable for short periods of time. Unfortunately, they are often provisioned as if an outage would be catastrophic. To avoid such wasteful overprovisioning, it is important for application developers and business owners to work with IT departments to define appropriate levels of performance, resiliency, and recovery for each of their applications.

Furthermore, if applications are designed with energy and resource efficiencies as key criteria, reliability is likely to be better than it would be when using traditional IT resource provisioning practices, which are prone to uncertainty and error. Applications that are able to dynamically report their performance and adjust to constraints should be much easier to keep in compliance with their service level agreement (SLA) compared to applications whose performance and availabiity requirements are only expressed on paper. Applications designed with SLA management in mind can also help eliminate redundant and expensive layers of resiliency that often plague traditional high-availability designs.

End User Computing Environments

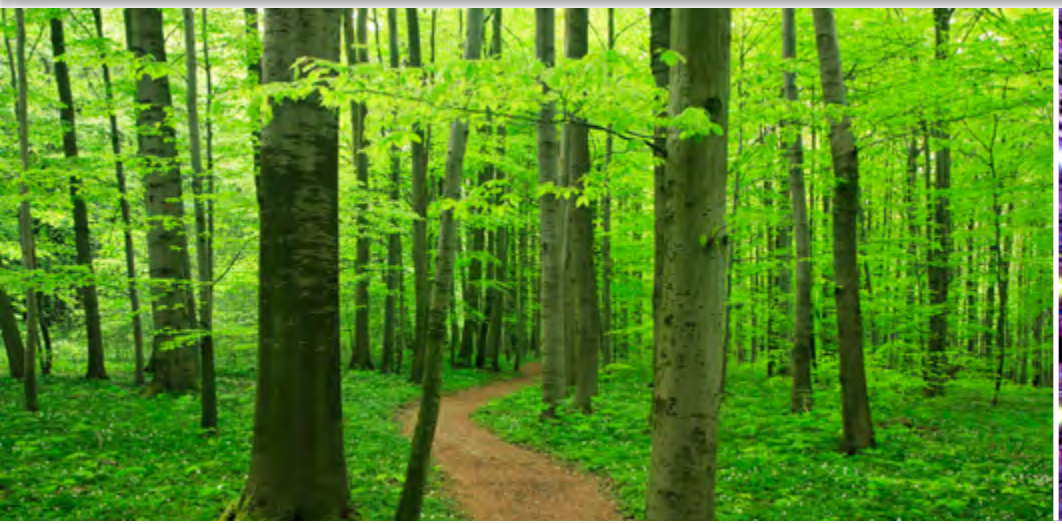
Lack of IT department control over end user computing environments can lead to a significant amount of wasted energy and higher-than-necessary office building cooling costs.[27] Organizations with large numbers of PCs and monitors should transparently reflect the cost of space, cooling, and direct energy use by employee equipment through charge-backs to their business units.

IT departments can also ensure that employees choose energy-efficient PCs, make certain that power management capabilities are enabled and optimized, and require that decommissioned PCs are unplugged, returned, and disposed of properly. Organizations with large numbers of users who need computing capacity beyond their primary device should also consider using server-based computing technologies such as Microsoft Desktop Virtualization.[28]

It is particularly important that developers of PC applications ensure that applications support the power management capabilities of the hardware and operating system—for example, by allowing PCs to automatically sleep when idle but not while they are doing something useful. Applications that are "aware" of the current power source—battery or AC power—and react accordingly can also help improve the productivity of mobile users. Developers of Windows-based applications can find more information on these energy-smart development practices at the Windows Developer Center at http://microsoft.com/energysmart.

To Sleep or Not to Sleep?

Because the computer manufacturing process itself has a significant environmental impact, arbitrarily turning computers off when idle is not necessarily the best way to amortize this impact over the life of the equipment. A better strategy may be to have idle computers perform useful computation instead, assuming excess power is available on the grid—particularly from renewable sources whose energy would otherwise be wasted. Mechanisms to enable this type of use could be integrated into public "volunteer computing"[29] platforms and could conceivably involve some form of reimbursement to the energy bill payer for the electricity consumed.



A Vision of IT Energy Efficiency

IT decision makers can better meet the needs of their organization and respond to demands for new IT services more quickly and cost effectively if "energy efficiency by design" becomes a fundamental IT tenet. Imagine this future scenario:

IT departments in large organizations have achieved significant operational flexibility and energy efficiency by running IT as a utility, using public cloud computing platforms for line-of-business applications and commodity services such as email, as well as their own private clouds for applications that must remain on premises for compliance or technical reasons. Smaller organizations mostly use applications run on a public cloud, having retired all of their servers except those running legacy applications that must run on premises.

The vast majority of new applications are developed and deployed directly on public and private clouds. Most legacy applications that were not designed with the cloud in mind have been migrated to private or public clouds as virtual machines. Departments outside of IT rent cloud computation-and-storage capacity through the IT department and are not allowed to buy servers to be housed in the data center. The IT department helps determine where the application should reside (private or public cloud) based on technical and regulatory constraints, and it passes the operating costs to the application owner.

Owners of applications that are deployed in a private cloud provide periodic usage forecasts to the IT department to help determine the quantity of IT hardware needed to adequately support demand. Usage trend reporting and instrumentation in new applications makes this forecasting easier.

Because application owners pay for computing resources on a frequent (e.g., hourly) basis, they have an incentive to ensure that

their applications can be dynamically scaled based on demand and to implement throttling mechanisms for applications with unpredictable demand. Many applications provide mechanisms to postpone noncritical work to provide additional "virtual" capacity for critical application services if there is a shortfall in IT resources or if power is constrained. Applications designed specifically for the private cloud can extend into the public cloud (in a process known as cloud bursting) if additional capacity is required.

Many of the new applications are also designed to be resilient enough to survive server and data center failures without the need for expensive clustering or other failover technologies. As a result, the overall utilization of powered-on servers dramatically increases.

Because the IT department buys, owns, and controls all of the IT hardware within the organization, it determines when the hardware will be deployed, powered on and off, refreshed, or decommissioned to maximize energy efficiency and productivity. Server configurations are right-sized and balanced to optimize utilization by the application portfolio. With the aid of software and hardware power management technologies, computers continue to improve their energy use by more closely scaling it with IT utilization. Excess capacity, particularly when initially deployed, can be temporarily turned off until it is needed, but it is ideally made available on a spot market for computation cycles when it is not being used.

The data centers themselves are constructed with energy-efficient components and a minimal environmental footprint. They require little protection from the elements because they are built with modular, weatherized components, which cost dramatically less in terms of materials (such as concrete and steel) and construction costs. Where possible, waste heat from the data centers is reused to preheat water for other commercial or residential purposes.

The organization's client computing infrastructure is similarly energy efficient. The power settings of desktop PCs are centrally managed; the PCs automatically sleep when idle but can be woken (even remotely) by the end user or system administrator. IT pros ensure that applications running on mobile and desktop PCs are energy-smart and don't keep the devices awake when they are not in use. Users who temporarily need additional computers can "check out" virtual machines running on servers to avoid buying an additional PC that will likely be underutilized over the long term.

Corporate IT hardware purchasing policies require that all hardware—servers, clients, displays, storage, networking, and peripherals—meets strict energy efficiency and IT resource consumption criteria and is designed with the environment in mind. This includes a strategy for effective and responsible reuse and recycling of unneeded equipment.

IT departments that operate in this way can be dramatically more responsive to their organization's needs and significantly reduce the amount spent on IT across the organization. If implemented at scale across the countless IT organizations worldwide, these actions would enable IT use to grow considerably while conserving an enormous amount of energy, raw materials, and water and significantly reducing IT-associated carbon emissions and other pollution. These practices could lead to a doubling or tripling of the average server utilization rate and could cut data center energy use by more than half—essentially flattening the current growth rate.

Immediate Opportunities

While the vision described on the previous several pages may seem aspirational, many of the enabling technologies are already available and the rest are on many IT vendors' roadmaps. Perhaps the greatest barrier to organizations adopting this vision is inertia. Fortunately, there are many opportunities to improve energy efficiency at all layers of the IT stack that do not require a wholesale re-engineering of how an organization runs IT. While it is important to keep the end goal in mind—operating IT as a utility—IT energy efficiency can be improved significantly through the use of components that require less power to run, by managing components so they do not need to run continuously at the same power levels, and by reducing the amount of hardware needed to do the job. Figure 2, on the next page, illustrates the various IT layers and the opportunities to improve energy efficiency at each layer.

The layers in the IT stack are interdependent, and their boundaries are not entirely distinct. For instance, operating system power management can be a coordinated effort between the Silicon and Operating System layers and can even include the Hardware Package layer, using firmware that enables cooperative processor power management. For an organization with multiple data centers, the Management Infrastructure layer will likely extend into the Building layer to allow for management of applications across multiple data centers.

Opportunities to Improve the Energy Efficiency of IT Operations

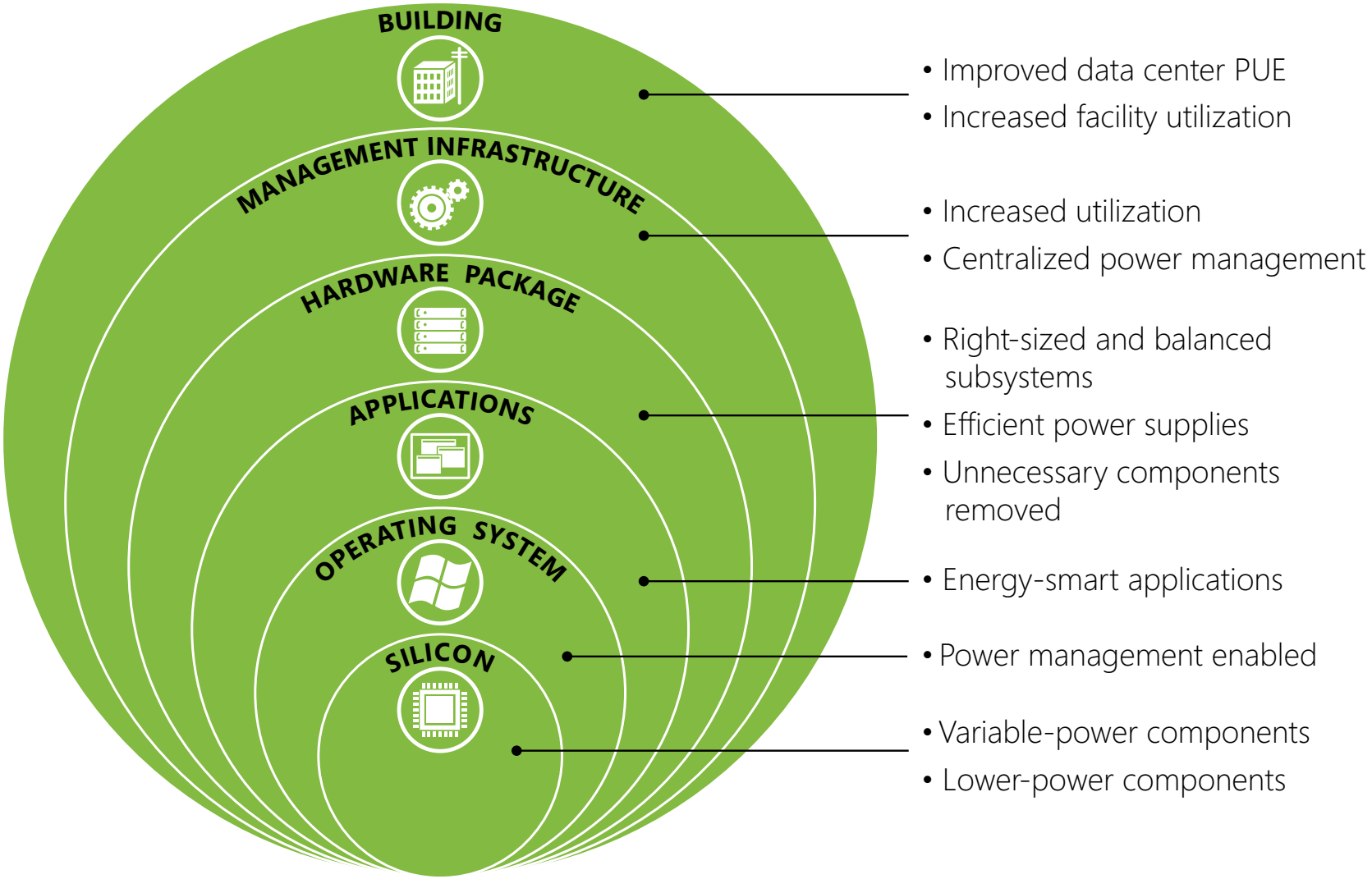


Figure 2.

Here are some IT energy efficiency opportunities to consider pursuing at the various layers.

IT Layer	Opportunity	Description
Silicon 	**Use variable-power components**	Certain components, such as the CPU and hard disk drives, can lower their power needs when less busy or idle, typically in conjunction with the operating system.
	Use lower-power components	Certain components, such as "green" RAM and disk drives, can use less power at normal operational loads through lower voltage or different designs (e.g., solid-state drives instead of hard disk drives).
Operating System 	**Enable operating system power management**	By understanding usage patterns, the operating system can help the hardware reduce its energy use.
Applications 	**Ensure use of energy-smart applications that cooperate with operating system power management**	Applications that are designed to work well with power management can ensure that servers and PCs are able to save energy when idle and that user productivity is not affected by displays or systems powering off when critical tasks are running.
	Design server applications to dynamically scale and be resilient against sudden hardware outages	Server applications that are designed to use IT resources dynamically and be tolerant to sudden equipment failure can dramatically improve server utilization by reducing the number of servers or virtual machines assigned to a given application. (See the earlier section titled "Applications: The Missing Link" for more information.
	Provide mechanisms to postpone noncritical tasks	Applications should be able to suspend or postpone noncritical operations when resources (IT resources or electric power) are constrained.
Hardware Package 	**Right-size and balance hardware subsystems**	Based on examination of an application's performance characteristics, hardware can be configured to ensure that subsystems (CPU/RAM/disk/network) are neither "starved" due to bottlenecks in other systems nor overbuilt and mostly idle. This balancing activity can significantly reduce the system's overall power draw and reduce costs while potentially improving performance.[30]

IT Layer	Opportunity	Description
Hardware Package (Continued) 	**Use efficient power supplies**	For systems such as servers that are on nearly 24/7, an efficient power supply can significantly reduce the amount of energy consumed and more than pay for itself in energy savings. For instance, a server that requires a near-constant 500 watts of DC power will consume about 1100 fewer kWh per year using an 85% efficient power supply than one that is 70% efficient ($110/year savings at 10 cents per kWh).
	Remove unnecessary components	Most computer hardware components draw power whether they are used or not. Configuring hardware to contain only components that are needed (e.g., no sound cards in servers) can reduce an organization's energy bill, particularly at scale (i.e., when you have many servers with the same configuration). Similarly, server costs (and associated environmental impact) can be reduced by eliminating unnecessary "cosmetic" plastics or even sheet metals.
Management Infrastructure 	**Improve server utilization**	By using technologies such as virtualization and virtual machine migration, servers can be kept at higher rates of utilization, reducing the amount of hardware needed in the data center. Application frameworks designed to abstract individual operating system instances from the application can also help improve server utilization.
	Monitor and control power management across PCs	Modern operating systems ship with power management enabled, but users can easily reduce this feature's effectiveness by increasing timeouts or by disabling it altogether. By deploying a centralized power management solution, IT departments can ensure that power management is used appropriately and monitor its effectiveness.
	Reduce data footprint	In addition to storage virtualization that allows storage hardware to be shared among many systems, storage management software can help significantly curb the growth of storage needs (and the associated energy consumption) through techniques such as data de-duplication, compression, and archiving.
Building 	**Monitor and improve data center resource usage effectiveness (PUE/CUE/WUE)**	By measuring energy used by the IT equipment in the data center, IT departments can calculate the data center's effectiveness at using power for its intended purpose (i.e., running servers) rather than for powering and cooling the data center itself. Metrics are also being developed to measure the ratio of water and carbon emissions to power consumed by the IT infrastructure.
	Increase facility utilization	As server workloads are consolidated onto fewer, more energy-efficient servers, it is important to understand the power and thermal load limits of the data center's cooling and power systems. Empty space in the data center is not only inefficient from a PUE perspective but also a wasted asset.

Principles and Practices

Taking into account all of the opportunities for increasing IT energy efficiency, how can an organization best approach the challenge of making its IT operations more sustainable? Here are the key principles for a successful transition to operating IT as a utility:

- Improving IT energy efficiency will significantly reduce the financial and environmental costs of running IT and make IT more responsive to business demands.
- Increasing IT resource utilization is the most effective way to improve IT energy efficiency.
- Central control of the IT infrastructure (local or cloud-based) is a key requirement for increasing IT resource utilization.
- Applications designed to optimize IT resource utilization are necessary to maximize the effectiveness of a centrally controlled IT infrastructure.

The following practices support the aforementioned principles:

- Continuously measure, report, and set goals for IT resource utilization.
- Ensure that the costs of deploying and operating applications on dedicated hardware are fully reflected in the cost charged to the application owner—including energy use, allocated power, and data center space.
- Centralize the budget for server hardware capital expenditures, and decentralize the budget for server hardware operations.
- Use centralized computation-and-storage services (based in private and/or public clouds), and rent out capacity to business units to recover costs.
- Adopt application designs that enable dynamic resource allocation, cloud bursting, controlled deferral of noncritical work, and workload power efficiency.
- Implement power management on the desktop, ensure that applications function correctly with power management enabled, and enact policies that mandate the procurement of energy-efficient hardware.

Setting the Right Goals

Energy efficiency changes in one layer of the IT stack can affect energy efficiency in other layers. For instance, improving the energy efficiency of servers in a data center by 50% can also save energy at the building layer because less energy is required for power distribution and cooling. If you reduce server power consumption and you don't add new servers to consume the power difference, you could end up actually making the PUE worse because the cooling systems in many data centers are sized to extract a specific range of heat. This might seem like a bad thing, but if you are producing at least the same amount of useful work with less energy, it's worth the sacrifice.

Seizing the Opportunity

The importance of embracing IT energy efficiency has never been more urgent than it is today. The integration of IT into almost every facet of business and society is driving exponential demand that will strain many organizations' finances and IT capabilities to the limit. IT energy efficiency and, in particular, increasing IT resource utilization from today's typically low levels, can offer substantial respite from these challenges and lead to more nimble, competitive organizations.

Opportunities to improve IT energy efficiency are numerous and increasing, but incentives and motivations are sometimes lacking. C-level decision makers should consider the significant cost savings and productivity opportunities that improvements in IT energy efficiency offer, as well as empower their IT departments to aggressively pursue them.

Organizations that wish to ensure that their IT capabilities are not constrained over the long term should begin the transition to operating IT as a utility, leveraging a mixture of private and public cloud technologies. Doing so could require changing how IT is funded and operated within the organization.

Business units that are used to having extreme levels of resiliency for their applications should evaluate their perspective on what is truly necessary, and whether the alternative represents better value. Application developers who traditionally design applications assuming unlimited IT resources will need to ensure that their applications can scale dynamically with load and respond to constraints. Applications that are easier to manage to a sensible SLA will likely have better performance and be more economical and reliable than those designed and operated without regard for energy efficiency or resource utilization.

The imperative for embracing IT energy efficiency extends far beyond the immediate needs or concerns of any individual organization. More effectively utilizing IT equipment can significantly reduce the growing volume of unrecycled e-waste ending up in developing countries, with its attendant environmental and health risks.

Regulations on pollution from traditional sources of electricity are already increasing energy costs for data center operators in certain regions, and it may be simply a matter of time before the effects are felt by nearly every data center operator. While IT energy efficiency is a critical endeavor, operators of data centers powered primarily by environmentally harmful fuels may want to consider working with their electricity suppliers to source pollution-free renewable energy, if available, or investigate other instruments such as offsets that mitigate their consumption of pollution-based electricity sources.

Similarly, scarce water supplies are likely to threaten the cooling capabilities of many data centers in the future. Like electricity, increased costs and regulations on water, e-waste, and other raw materials will substantially impact the bottom line of many data centers over time. Embracing IT energy efficiency and improving utilization may be the only ways to ensure the viability of many IT operations.

Some people have expressed concern that advances in IT energy efficiency may, paradoxically, increase the environmental footprint of IT services. As the cost of computation, data storage, and network

bandwidth drops through performance and energy efficiency improvements, IT services and products might become even more widely used than already projected. However, if these more powerful and efficient computers are used effectively, it is feasible that much of the additional demand for IT services could be satisfied without a dramatic increase in the installed base of servers and data centers.

The opportunity and the imperative are clear. The only question is whether organizations will embrace IT energy efficiency and reap the rewards before it's too late. So, will you?

Endnotes

[1] http://newsroom.cisco.com/dlls/2010/ekits/Cisco_Connected_World_Report_PartIII.pdf ("Virtualized Servers: Production Environment," p. 13)

[2] www.infoworld.com/t/server-virtualization/what-you-missed-server-virtualization-has-stalled-despite-the-hype-901

[3] http://arstechnica.com/hardware/news/2008/05/study-recommends-datacenters-go-green-adopt-power-metric.ars

[4] www.datacenterknowledge.com/archives/2010/04/09/kundra-fed-data-centers-7-percent-utilized

[5] www.1e.com/contenthub/doc_download.aspx?id=44674081

[6] This is based on the assumption that a PC is on 24/7 but used less than 40 hours a week.

[7] This energy productivity metric is based on a data center with a PUE of 2.0, an average 7% server utilization, and 65% efficient power supplies, where energy productivity = server power * PSE efficiency * utilization / (server power * PUE). It is similar to the Compute Power Efficiency (CPE) metric from the white paper by C.L. Belady and C.G. Malone, "Metrics and an Infrastructure Model to Evaluate Data Center Efficiency," Proc. of the ASME 2007 InterPACK Conf., July 2007. The difference between this energy productivity metric and CPE is that CPE does not factor in power supply efficiency.

[8] There are obviously upper limits to the level of efficiency or utilization that can be achieved in any system; a percentage of this waste will be unavoidable. However, unavoidable waste can be reduced as technologies and operational practices evolve.

[9] www.thegreengrid.org/sitecore/content/Global/Content/white-papers/The-Green-Grid-Data-Center-Power-Efficiency-Metrics-PUE-and-DCiE.aspx

[10] 50% server utilization is the efficiency equivalent of a data center with a PUE of 2.0. If the baseline for server utilization is 50%, then energy productivity starts at 16% and climbs to about 27% when utilization improves to 83% or PUE reaches 1.2. In this case, improvements in either PUE or server utilization would have similar effects on energy productivity and follow the same trajectory.

[11] http://ec.europa.eu/environment/waste/weee/index_en.htm

[12] http://svtc.org/wp-content/uploads/greenpeace.pdf

[13] Energy Star aims to recognize the top 25% of models in a particular product category in terms of energy efficiency.

[14] www.epeat.net

[15] www.microsoft.com/Environment/products-and-solutions/cloud-computing.aspx

[16] www.ukcrc.co.uk/crc/legislation.htm

[17] 6,000 MWh/total hours per year (8,766) x 1,000 = 684 KW constant power draw. 200 W server x 1.7 PUE = 340 W to power and cool server. 684,000 W/340 = 2011 servers.

[18] http://ec.europa.eu/environment/waste/weee/index_en.htm

[19] http://en.wikipedia.org/wiki/ROHS

[20] www.computerweekly.com/Articles/2010/03/17/240633/How-businesses-can-prevent-their-old-IT-being-dumped-in-developing.htm

[21] http://247wallst.com/2010/07/07/record-heat-stresses-northeast-power-grid

[22] Forrester Research, "Market Update: The State Of Green IT Adoption, Q2 2010."

[23] http://newsroom.cdw.com/features/feature-11-08-10.html

[24] http://research.microsoft.com/en-us/projects/marlowe

[25] http://research.microsoft.com/en-us/projects/orleans

[26] Some data center operators are doing this today and relying on their backup generators, which have to provide power for the entire data center, even though a significant portion of the data center's workload could be temporarily halted to save money on fuel.

[27] A PC that consumes about 65 watts when idle generates heat output equivalent to that of a 70 kg person at rest.

[28] www.microsoft.com/virtualization/en/us/products-desktop.aspx

[29] http://boinc.berkeley.edu/trac/wiki/VolunteerComputing

[30] www.globalfoundationservices.com/blogs/documents/Rightsizing_WhitePaperFINALDecember09-2.pdf

Acknowledgments

We wish to thank the following people for their invaluable contributions, feedback, and editing on this white paper (Microsoft employees unless designated). Any omissions are purely unintentional.

Ori Amiga, Nancy Anderson, Christian Belady, Rob Bernard, Dileep Bhandarkar, David Bills, Ina Chang (Katz Communications Group), Sacha Dawes , Pierre Del Forge (Natural Resources Defense Council), Nic Delaye (PwC), Tom Harpel, Josh Henretig, Ulrich Homann, Dean Katz (Katz Communications Group), Jonathan Koomey (Stanford University), Haris Majeed, Christopher McCarron, Chris Mines (Forrester Research), Anil Nori, Barry O'Flynn (Mainstream Renewables), Alfredo Pizzirani, Shilpa Ranganathan, John Shaw (Mainstream Renewables), Kim Shearer, Shane Snipes (The Eleven Agency), Amaya Souarez, Jorgen Thelin, Marian Wachter (Katz Communications Group), CJ Williams, and Kathryn Willson.

Microsoft®