ISSUE PAPER

AUGUST 2014
IP:14-08-A

Data Center Efficiency Assessment

Scaling Up Energy Efficiency Across the Data Center Industry: Evaluating Key Drivers and Barriers

NRDC
The Earth's Best Defense

Anthesis

Acknowledgments

Project managed by:
Pierre Delforge (NRDC)
Josh Whitney (Anthesis)

Research performed by:
Anthesis

Report written by:
Josh Whitney (Anthesis)
Pierre Delforge (NRDC)

Reviews by:
Eric Masanet (Northwestern University)
George Peridas (NRDC)
Maria Stamas (NRDC)
Nick Jimenez (NRDC)
Pat Remick (NRDC)
John Clinger (ICF International, ENERGY STAR™ Servers Program)
Simon Mui (NRDC)
Ted Brown (Seattle City Light)

Important technical and substantive contributions were made by:
Mark Aggar (Microsoft)
Rhonda Ascierto (451 Research)
Chris Berry (Rocky Mountain Institute)
Austin Brown (NREL)
Braulio Calle (Unlimited Conferencing)
Brian Caslin (BlackRock)
Joyce Dickerson (Google)
Patrick Flynn (IO Data Centers)
Steve Knipple (EasyStreet Data Centers)
Mark Monroe (DLB Associates)
Nicole Peil-Moelter (Akamai)
John Pflueger (Dell)
Aaron Rallo (TSO Logic)
Winston Saunders (Intel)
Russel Sprunger (Intel)
John Tuccillo (The Green Grid)
Pitt Turner (Uptime Institute)
Otto Van Geet (NREL)
Bill Weihl (Facebook)

About NRDC

The Natural Resources Defense Council (NRDC) is an international nonprofit environmental organization with more than 1.4 million members and online activists. Since 1970, our lawyers, scientists, and other environmental specialists have worked to protect the world's natural resources, public health, and the environment. NRDC has offices in New York City, Washington, D.C., Los Angeles, San Francisco, Chicago, Bozeman, MT, and Beijing. Visit us at www.nrdc.org and follow us on Twitter @NRDC.

NRDC Director of Communications: Lisa Benenson
NRDC Deputy Director of Communications: Lisa Goffredi
NRDC Policy Publications Director: Alex Kennaugh
Design and Production: www.suerossi.com

TABLE OF CONTENTS

ACRONYMS AND DEFINITIONS

Colocation/Multi-tenant Data Centers	Shared data centers where customers lease space, power, Internet connectivity, cooling and security services to run their computing equipment rather than managing their own data center
CPU	Central Processing Unit
DCIM	Data Center Infrastructure Management
ICT	Information and Communications Technologies
IT	Information Technology
PUE	Power Usage Effectiveness
SMO	Small and Medium-Sized Organization
UPS	Uninterruptible Power Supply
Utilization	Ratio of processing load on a server relative to its maximum processing capacity
Virtualization	A technique that allows the consolidation of workloads from underutilized servers onto fewer servers

EXECUTIVE SUMMARY

Data centers are the backbone of the modern economy, from the server rooms that power small- to medium-sized organizations, to the enterprise data centers that support American corporations, to the server farms that run cloud computing services hosted by Amazon, Facebook, Google, and others. However, the explosion of digital content, big data, e-commerce, and Internet traffic is also making data centers one of the fastest-growing users of electricity in developed countries, and one of the key drivers in the construction of new power plants in the United States.

While most media and public attention focuses on the largest data centers that power so-called cloud computing operations—companies that provide web-based and other Internet services to consumers and businesses—these hyper-scale cloud computing data centers represent only a small fraction of data center energy consumption in the United States.

As NRDC initially found in its groundbreaking 2012 analysis, *Is Cloud Computing Always Greener? Finding the Most Energy and Carbon Efficient Information Technology Solutions for Small- and Medium-Sized Organizations,* smaller server rooms and closets are responsible for about half of all U.S. server electricity consumption—but 50 percent of that is wasted due to lack of awareness and incentives to make them more efficient. There remains a critical need for action, including developing utility incentive programs to reduce waste in the massive amounts of electricity used by data centers small and large.

In 2013, U.S. data centers consumed an estimated 91 billion kilowatt-hours of electricity. This is the equivalent annual output of 34 large (500-megawatt) coal-fired power plants, enough electricity to power all the households in New York City twice over. Data center electricity consumption is projected to increase to roughly 140 billion kilowatt-hours annually by 2020, the equivalent annual output of 50 power plants, costing American businesses $13 billion per year in electricity bills and causing the emission of nearly 150 million metric tons of carbon pollution annually.[1]

If just half of the technical savings potential for data center efficiency that we identify in this report were realized (to take into account the market barriers discussed in this report), electricity consumption in U.S. data centers could be cut by as much as 40 percent. In 2014, this represents a **savings of 39 billion kilowatt-hours annually**, equivalent to the annual electricity consumption of all the households in the state of Michigan. **Such improvement would save U.S. businesses $3.8 billion a year.**

There has been significant progress in data center efficiency over the past decade, with shining examples of ultra-efficient server farms run by the likes of Google and Facebook. But how efficient, really, is the typical data center in the United States? What are the key opportunities for further efficiency gains, and what are the main barriers to capturing these opportunities? Further, while projections of overall industry growth and its energy and carbon impacts remain elusive, even less is known about the makeup of the sector by data center type and the impact this may have on future carbon emissions. These are the questions that NRDC and Anthesis set out to answer in this extensive survey by interviewing more than 30 industry stakeholders and experts and reviewing the latest industry literature.

Much of the progress in data center efficiency over the past five years has occurred in the area of facility and equipment efficiency. However, little progress has been achieved in server operation efficiency, particularly in terms of server utilization. In addition, progress remains uneven across the different segments of the data center market. This study therefore focuses on assessing the current situation, the opportunities and the barriers related to server utilization efficiency, and the ways in which these opportunities and barriers vary across different segments of the data center market, particularly with regard to the multi-tenant data center business model.

This study found that while the largest public-facing companies providing cloud computing services typically run their data centers very efficiently, progress on energy efficiency has slowed in other types of data centers. Most significantly, servers are being used very inefficiently, consuming power 24/7 while doing little work most of the time. This is due to a variety of factors, such as these:

- Peak provisioning: Data center operators install enough equipment to handle peak annual load, and then some, but do not power down unused equipment during the majority of the time when it is not needed.
- Low deployment of virtualization technology (which allows the consolidation of workloads onto fewer servers) across the entire server fleet.
- "Comatose" servers: A large number of servers that are no longer being used still gulp energy 24/7 because no one is decommissioning them or is even aware that they're no longer used.

Multi-tenant data centers are shared data centers where customers lease space and power to run their computing equipment rather than managing their own data center. This creates a potential split incentive situation, where the decisions made by customers on the efficiency of their information technology (IT) equipment have no direct impact on their bills, removing any financial motivation to improve energy efficiency.

Split incentives, both internally within organizations and between multi-tenant data center providers and customers, remain one of the biggest reasons why efficiency best practices are not being implemented at scale across the data center industry. Only 20 percent of organizations' IT departments pay the data center power bill, a statistic that has not changed in five years. While this issue has been largely solved among hyper-scale cloud service providers, contractual relationships between colocation providers and their customers compound the split incentive challenge, with the data center owner paying the power bill, the tenants buying power blocks, and their IT purchasers separately specifying equipment.

The technical solutions, such as virtualization, server power management, and single responsibility for IT and facility costs, are well known but are not applied systematically. This report recommends system-level policy actions to create the conditions for best-practice efficiency behaviors across the data center market. One of these systemic levers is the adoption of a simple CPU utilization metric that would provide internal and external visibility to the IT efficiency of data centers, creating management and market incentives for operators to optimize the utilization of their IT assets. Others include public disclosure of data center energy and carbon performance, and the development of templates for green colocation and other multi-tenant data center contracts.

DATA CENTER EFFICIENCY

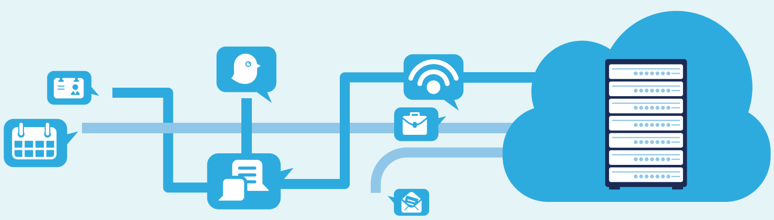
12 million computer servers in nearly **3 million** data centers deliver all U.S. online activities.
Email, social media, business, etc.

They gulp enough electricity to power all of NYC's households for 2 years.

That's equivalent to the output and pollution of

34 coal-fired power plants.

Many big "**cloud**" computer server farms do a great job on efficiency, but represent **less than 5% of data centers' energy use**. The other 95%—small, medium, corporate and multi-tenant operations—are much less efficient on average.



A typical data center wastes large amounts of energy powering equipment doing little or no work. **The average server operates at only 12-18% of capacity!**

Action is needed to accelerate adoption of energy efficiency best-practices.

Achieving just half of technologically feasible savings could cut electric use by 40% and save U.S. businesses $3.8 billion annually.

NRDC
The Earth's Best Defense

Learn More at
nrdc.org/energy/data-center-efficiency-assessment.asp

1. INTRODUCTION

1.1 OBJECTIVE

In light of the continued, rapid growth of the data center industry, including the emergence of new business models broadly defined as cloud computing, NRDC retained Anthesis to conduct a study to assess progress on energy efficiency in the data center industry. The study focused on three key data center issues: the level of utilization of IT equipment, the impact of and potential for efficiency opportunities in multi-tenant data centers, and the degree to which the evolution of the industry's technology and delivery model is aligning incentives to further drive energy efficiency.

Server utilization represents processing load on the server relative to maximum server capacity, like the number of passengers on a bus relative to the total number of seats. This is a key factor in data center efficiency because a server's efficiency drops dramatically as its utilization level decreases.

Existing research suggests that low server utilization remains one of the largest opportunities for energy savings in data centers. However, the optimization of server utilization faces several barriers that are hindering progress. Additionally, the multi-tenant data center business model, by which customers lease space and power to run their computing equipment rather than managing their own data center, is becoming increasingly popular. This presents split incentive issues, with customers having little or no inducement to implement energy efficiency best practices, thus limiting the deployment of energy efficiency opportunities.

This study is based on primary research conducted by Anthesis, including interviews with more than 30 industry stakeholders and experts representing the full spectrum of the data center industry and its organizational roles. Secondary research included a review of published articles, market research, and consideration of best practice standards and emerging policies. Anthesis identified secondary information sources from leading experts, organizations, and data center industry associations to help inform its analysis of key barriers, challenges, and recommendations concerning server utilization and multi-tenant data centers. These methods provided relevant background information for this report. Expert interviews were used as a vehicle to vet, share, and collaborate around the concepts identified and proposed in this report.

1.2 BACKGROUND

The data center industry is rapidly evolving to offer a complex blend of services, business models, and infrastructure deployments as consumers and businesses demand 24/7 connectivity, increased flexibility of resources, and lower costs for information technology (IT) services. While small enterprise server rooms remain a large portion of the market, a variety of data center options have emerged, from private and hybrid data centers to colocation and multi-tenant data center hosting providers. This evolution is also marked by the growth in cloud computing services that are hosted and operated largely by companies such as Amazon, Apple, eBay, Facebook, Google, Microsoft, and Salesforce.com.

The Internet has created myriad new opportunities for society and the economy, but its backbone, the data center industry, represents a significant environmental burden due to its energy consumption. If the worldwide Internet were a country, it would be the 12th-largest consumer of electricity in the world, somewhere between Spain and Italy.[2] This represents roughly 1.1 to 1.5 percent of global electricity use (as of 2010) and the greenhouse gases generated annually by 70 to 90 large (500-megawatt) coal-fired power plants.[3] While this figure indicates a reduction in the pace of growth from previous estimates, the continued expansion of the industry means that the energy use of data centers, and the associated emissions of greenhouse gases (GHGs, often referred to as carbon emissions or CO_2e) and other air pollutants, will continue to grow.[4] Industry experts struggle to put numbers to the potential growth rate, but one of the more widely recognized studies, the SMARTer 2020 report, shows global data center emissions will grow 7 percent year-on-year through 2020.[5] Although an increasing number of studies suggest that information services delivered over the Internet have the potential to lead to significant reductions in broader societal carbon emissions, this potential cannot be an excuse for the very real and increasingly substantial climate and environmental impact of data centers themselves. Further, as the market transitions to a variety of data center types and models, the variability in energy efficiency may significantly impact the sector's future carbon emissions impact.

Over the past decade, the industry has made great strides in identifying the fundamental challenges of data center efficiency, with member-based organizations like the Green Grid and the Uptime Institute, as well as federal government agencies like the U.S. Department of Energy and the Environmental Protection Agency, supporting the implementation of energy efficiency efforts by issuing guidance, identifying best practices, and developing standards. An initial focus has been the efficiency of the cooling infrastructure of data center facilities, which typically require one to two times the energy used to power the IT equipment itself.

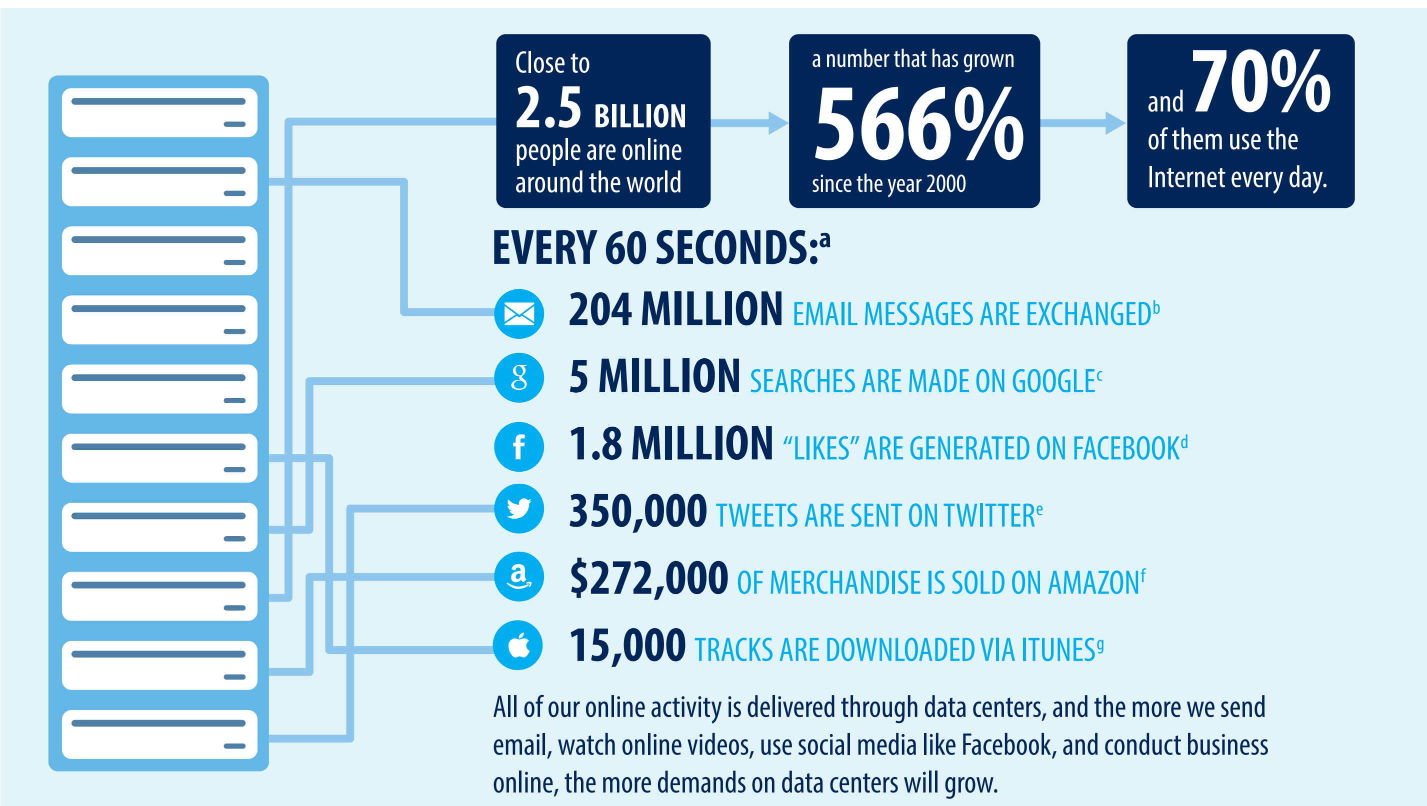
Motivated by cost, publicity, and pressure from environmental organizations, and incentivized because they own and operate their data centers, the largest consumer-facing companies like Google, Facebook, eBay, Microsoft, and Apple have largely solved the technical efficiency problems posed by such facilities. They are highly efficient due to their economies of scale, diversity, and aggregation of users; flexibility of operations; and ease of sidestepping organizational constraints.[6] But these companies' data center facilities represent only an estimated 5 to 7 percent of the total installed base of servers globally.[7] Most segments of the industry have yet to adopt best practices and are failing to capture the majority of efficiency opportunities. These segments range from small server rooms to corporate-owned (enterprise) data centers to the rapidly growing multi-tenant data center segment. The latter, in particular, is seeing rapid year-to-year growth rates of 18 to 20 percent and is subject to a number of unique challenges that have made the deployment of energy efficiency especially difficult.[8,9]

1.3 THE DATA CENTER INDUSTRY

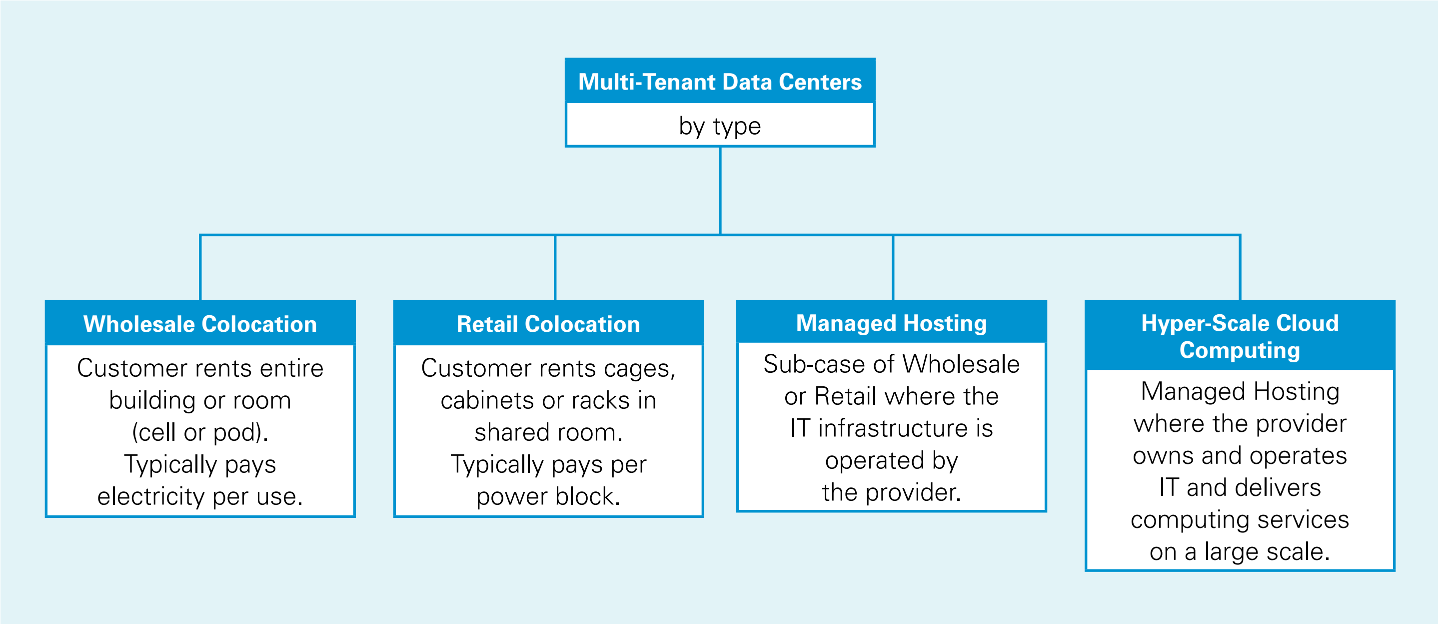
Data centers are facilities that contain information technology equipment including computer servers used for data processing, data storage devices, and networking devices. Data centers also contain infrastructure equipment, which typically consists of specialized power conversion and backup hardware (to ensure a reliable electricity source) and environmental control equipment (to maintain acceptable temperature and humidity conditions).[10]

While small, medium, and large enterprises can (and do) own their own data centers, consisting of everything from a small server closet to a few thousand square feet of dedicated server room space, a new IT delivery method has emerged in the past 10 years to provide these services to multiple customers in a shared facility. This new model is the multi-tenant data center.

Multi-tenant data centers provide data center services to customers on a lease basis. They also provide space and/or services to individual enterprises that place and manage their own equipment while the provider manages the cooling and facility infrastructure. Cloud, managed services, and shared hosting providers, while falling into this multi-tenant data center category, provide three layers of service: infrastructure, platform, and software-as-a-service. The largest of these companies, often referred to as hyper-scale cloud computing



a As of 2013
b http://www.domo.com/learn/infographic-data-never-sleeps
c http://www.statisticbrain.com/google-searches/
d http://gizmodo.com/5937143/what-facebook-deals-with-everyday-27-billion-likes-300-million-photos-uploaded-and-500-terabytes-of-data
e http://www.internetlivestats.com/twitter-statistics/
f http://www.domo.com/learn/infographic-data-never-sleeps
g http://www.billboard.com/biz/articles/news/1538108/itunes-crosses-25-billion-songs-sold-now-sells-21-million-songs-a-day



providers, offer services ranging from outsourcing of all infrastructure and applications to a hybrid model that enables organizations to keep their computing resources in-house and leverage cloud computing resources only when needed.

Within the multi-tenant data center industry, it is important to differentiate among a number of delivery and service models, including the following:[11] (A full description of each segment is provided in Appendix 2.)

1. Wholesale colocation
2. Retail colocation
3. Managed hosting
4. Hyper-scale cloud computing

It is difficult to quantify the size of each delivery model in terms of its total percentage of the data center industry, but pervasive adoption has many believing that multi-tenant data centers will become the "new normal," as companies that currently own and operate data centers determine what can be incorporated into their existing technologies and services and what makes economic sense to outsource to the cloud.[12]

The last segment of the data center market is high-performance computing (HPC), which comprises data centers owned and operated by universities, research facilities, and national laboratories, such as the National Renewable Energy Lab and the Lawrence Berkeley National Lab.

Using a combination of sources, Figure 1 (on the following page) shows the estimated electricity consumption by each segment within the data center market, based on the number of installed servers and estimated infrastructure (IT and facility) electricity consumption (see Appendix 2 for more information). Small- and medium-sized server rooms continue to account for nearly half the electricity consumption of the market; they are typically very inefficient, and there are millions of them, found in the office buildings of both small and large organizations throughout the country. Enterprise/Corporate data centers account for roughly a quarter and multi-tenant data centers for nearly a fifth of data center energy consumption. Finally, while much focus has been on the energy use of hyper-scale cloud providers such as Google, Apple, and Facebook, their consumption is very small compared with that of the other segments, in part a testament to their ability to aggressively deploy energy efficiency measures for both facility and IT systems.

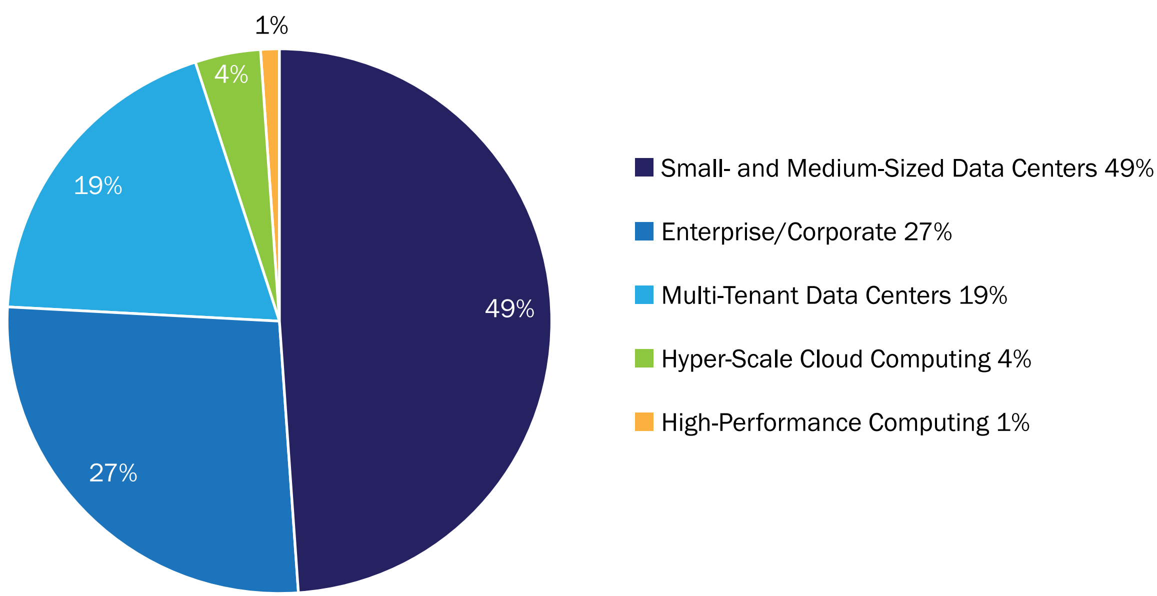
Given the lack of consistent and accurate data across all of these segments and the complexity in the assumptions used, these estimates contain uncertainty and will require further research and analysis. Nonetheless, for discussion purposes, the estimated market segmentation provides a new perspective on where energy consumption is occurring within this market and can be used to focus efforts on improving energy efficiency in the largest and fastest-growing segments.

Table 1: Estimated U.S. data center electricity consumption by market segment (2011)

Segment	Number of Servers (million)	Electricity Share	Total U.S. Data Center Electricity Use (billion kWh/y)
Small and Medium Server Rooms	4.9	49%	37.5
Enterprise/Corporate Data Centers	3.7	27%	20.5
Multi-Tenant Data Centers	2.7	19%	14.1
Hyper-Scale Cloud Computing	0.9	4%	3.3
High-Performance Computing	0.1	1%	1.0
Total (rounded)	**12.2**	**100%**	**76.4**

See Appendix 2 for source information

Figure 1: Estimated U.S. data center electricity consumption by market segment (2011)



See Appendix 2 for source information

2. DATA CENTER ENERGY EFFICIENCY CHALLENGES

Good progress has been made in implementing energy efficiency measures within the data center industry, and there are numerous shining examples of ultraefficient data center facilities. However, broad adoption of best practices has yet to take place across the sector, specifically within the three largest segments:

- Small server rooms of small- to medium-sized organizations
- Corporate/enterprise data centers
- Multi-tenant data centers

Many of the key, overarching barriers to efficiency already have been exhaustively identified and discussed by trade publications, but they remain largely present today and make achieving high-performing, cost-effective, and carbon-efficient data centers a challenge.

2.1 PROGRESS ON ENERGY EFFICIENCY IS SLOWING

Beyond the largest public-facing companies providing cloud services, progress on energy efficiency is slower than it needs to be. While rising energy costs are an incentive for greater efficiency, the pressure to keep up with technological developments and the rapid pace of operational change mean that energy efficiency is often a lower priority.

Data center operators have the opportunity to reduce the energy use of both their IT equipment and their facility's cooling infrastructure. These efficiency measures can reduce electricity bills (operational expenditure, or OpEx) and help operators defer or avoid expansion or construction of a new facility (capital expenditure, or CapEx).

However, beyond the largest public-facing companies providing cloud services, progress on energy efficiency seems to have slowed and may be considered a lower priority. According to the Uptime Institute's most recent (2013) survey of more than 1,000 global data center operators, only half of North American respondents said they considered efficiency to be very important. Following initial gains in power usage effectiveness (PUE), most operators consider a 1.65 PUE (the average for the survey) good enough, even though some hyper-scale cloud providers are consistently achieving PUEs below 1.1 in their data centers.[13] While hyperscale cloud providers often use custom IT and cooling equipment not readily available commercially, this provides a benchmark for the efficiency potential in other data centers, either through getting equipment vendors to adopt the higher efficiency technology from custom cloud solutions in their offerings or by moving their computing loads to the cloud.

In addition, a survey conducted by Digital Realty Trust in January 2013 indicated that only 20 percent of the 300 North American data center companies with revenues of at least $1 billion and/or more than 5,000 employees have a PUE below 2.0, with the average at 2.9.[14] Once data center operators have implemented low-hanging-fruit projects such as properly isolating hot and cold aisles, installing blanking panels in unused rack segments, and upgrading old power distribution equipment to more efficient models, improvements are more difficult and costlier to come by, resulting in longer paybacks that are harder to justify to management.[15] Also, it is challenging for cloud, wholesale, and multi-tenant providers to manage and optimize data center operations and capacity (power, space, cooling) in a complex, fast-paced environment where multiple customers may be utilizing the same application, server, platform, or infrastructure on a second-to-second, day-to-day basis.

Even though the increasing cost of energy and its limited availability in capacity-constrained facilities are considered core components of data center planning and strategy, efficient use of that energy is often a lower priority amid larger concerns for data center operators. Further, projects that may have a dramatic impact on efficiency are often not implemented because of the perceived risk that they might affect continuous operation or performance. Many managers have maintained their reputation by being extremely cautious in an industry where even a minor outage could threaten their employment.[16] As a result, once a facility is operational and populated with servers, the incentive to undergo major retrofits or to change operational behaviors is significantly diminished.

The tools historically used to manage the complex environment of a data center are now proving to be inefficient or inadequate for many. Most operators use a loose collection of tools that each address one or two specific operational elements, but there is little integration among all elements, making it difficult for management to solve

Power Usage Effectiveness (PUE)

A concept developed by the Green Grid and now widely considered the de facto industry standard for performance measurement, power usage efficiency is a ratio of the amount of energy a data center consumes relative to the amount its IT equipment uses. If no power losses were to occur and no additional power were required for cooling, the PUE would be 1.0. A PUE of 2.0 means that for every watt used by IT equipment, another watt is used by cooling, power distribution, and lighting equipment.

complex issues, such as how to maximize efficiency without compromising availability.

Data center infrastructure management (DCIM) systems are devised with the intention to transform the operational efficiency, reliability, and financial management of data center facilities by better aligning IT systems with facilities systems. DCIM helps managers track and analyze information about their data center's operational status, assets, and resource use (space, power, cooling, and so on). It also includes optional advanced features for planning, forecasting, optimization, and dynamic control. Some DCIM capabilities are also beginning to integrate or converge with IT management systems so that companies can better match the supply of data center resources with IT demand.

However, a confluence of factors is slowing DCIM adoption, which is used in approximately 15 to 40 percent of data centers in the United States today. Cost remains the largest barrier: In the Uptime Institute survey, more than 60 percent of respondents globally said the cost of DCIM was too high. Although DCIM is not necessarily expensive when compared with other enterprise management software systems, it is rarely planned for, and many data center operators do not have a large budget for software implementation. Also, because DCIM crosses the typical divide between IT and facilities departments within companies, there are questions over who should pay for it.[17] Finally, developing a clear path to return on investment remains a considerable challenge for DCIM vendors.[18]

Considering the well-publicized efforts within the data center industry to cut energy use, this overall stagnation on efficiency savings is concerning. There is clearly room for further infrastructure efficiency improvements in most data centers. Additionally, there are many data center efficiency opportunities that exist on the IT side. In fact, a recent study of the characteristics of low-carbon data centers and the factors that affect them showed that IT efficiency (which includes higher utilization and performance improvements as well as the purchase of efficient hardware) is the most important issue on which to focus when looking at overall energy and carbon efficiency. The two other factors, improving infrastructure efficiency (as measured by PUE) and locating data centers in areas where low-carbon electricity is available, are important but secondary to IT efficiency.[19]

2.2 SERVER UTILIZATION REMAINS LOW

Studies show that average server utilization remained static at 12 to 18 percent between 2006 and 2012. This underutilized equipment not only has a significant energy draw but also is a constraint on data center capaity.

After PUE, server utilization is perhaps one of the most common topics referenced in discussions of data center energy efficiency. Most servers in data centers, sometimes thousands of them, are on 24/7, waiting to receive data or transaction requests. Data center operators typically plan the number and processing capacity of servers to handle peak annual traffic, such as Black Friday sales; the rest of the time, most servers remain largely unused. Yet in this idle or nearly idle mode, they still consume substantial amounts of energy. Unfortunately, this situation remains largely unnoticed because utilization as a metric is rarely tracked and reported.

Even though the past decade has seen data center IT managers adopt server virtualization broadly (a technique to consolidate underutilized servers), average server utilization is still between 12 and 18 percent and has remained static from 2006 through 2012.[20,21,22] The limited industry focus, collecting, and reporting of this metric neglects its importance in determining and driving IT energy efficiency.

Although hyper-scale cloud providers can realize higher utilization rates (ranging from 40 to 70 percent), even they are not consistently achieving those rates.[23] New research from Google indicates that typical server clusters average anywhere from 10 to 50 percent utilization.[24]

There are many factors that directly contribute to low server utilization. They include:

1. vast over-provisioning of IT resources;
2. limited deployment of virtualization despite its broad penetration;
3. unused ("comatose") servers;
4. inherent limitations to high utilization levels outside of public cloud computing;
5. under-deployment of server power-management solutions;
6. procurement practices focused on initial cost instead of total cost of ownership; and
7. lack of a common, standardized server utilization metric.

Server utilization, which represents processing load on the server relative to maximum server capacity, like the number of passengers on a bus relative to the total number of seats, is a key factor in data center efficiency because server efficiency drops dramatically as its utilization level decreases. Servers still use 30 to 60 percent of their maximum power when utilized at 10 percent or lower capacity, doing little or no work. They can perform many times more transactions for a given amount of energy at high utilization than at low utilization, just as a bus gets many more miles per gallon per passenger with 30 passengers than with 5 passengers.

2.2.1 Vast over-provisioning of IT resources

"Peak provisioning," or planning for a capacity reserve to cover peak load events, generally results in suboptimal operating efficiency. Few operators implement power management and other techniques that can alleviate the inefficiencies associated with this approach.

Across each market segment, the over-allocation, or provisioning, of IT resources is leading to unnecessary expenditures in energy and poor utilization of equipment. When developing a power profile for a rack of equipment, an IT manager will often apply a 20 to 30 percent margin on the amount of power needed based on the equipment load to cover peak periods. While it is critical to ensure availability and uptime for these peak periods, it is unnecessary for the servers to be powered on continuously at times when the load is far from the projected peak. Also, peak capacity could be drawn from the cloud, using hybrid cloud solutions instead of being provisioned locally. Within the multi-tenant data center setting, many providers will add an additional 10 to 20 percent over peak requirements to ensure any one customer has sufficient power availability at all times.

Added together, the margins applied by the IT manager and the provider often result in nearly 50 percent extra power being allocated and, in the case of a multi-tenant customer, charged, regardless of actual use. This model of charging customers by "power block," irrespective of how much energy they use over time, has historically been a way for many multi-tenant providers to enhance profitability in their operations; it also represents a barrier to changing the pricing model around power, space, and appropriate resource allocation.

2.2.2 Limited deployment of virtualization despite its broad penetration

The technique of virtualization enables significant and cost-effective gains in server utilization and associated energy efficiency. While many data centers have implemented some level of virtualization, its overall deployment across all data centers remains much lower than it could be.

Few technologies have become a fundamental part of the data center as quickly as server virtualization. That is because the basic value proposition is so easy to grasp: When you run many virtual servers on a single physical server, you get a lot more out of your hardware, so you can invest in fewer physical servers to handle the same set of workloads. Reducing the number of servers yields indirect cost savings—less space to rent, less cooling to pay for, and, of course, lower power consumption. Even more compelling is virtualization's inherent agility: As workloads shift, you can spin up and spin down virtual servers with ease, scaling to meet new application demands on the fly.[25]

But while most organizations have implemented some level of virtualization, the overall share of virtualized servers remains much lower than it could be. Many research reports, and even virtualization software vendors, have published numbers indicating that although most enterprises have stuck their toe into the waters of virtualization, they have hardly plunged in. A 2010 survey by Prism Microsystems found that just 30 percent of production servers have been virtualized. More recent reports and surveys, released by the research firm Gartner and the virtualization software provider VMware in 2012, put overall adoption rates consistently in the 50 to 75 percent range for servers worldwide, although

Case study

A large corporation replaced 3,100 single-workload servers with 150 virtual hosts (physical servers hosting multiple virtual servers), achieving a $2.1 million annual reduction in electricity use and the recovery of $14 million in facility capital. The reclaimed power and cooling capacity allowed the company to avoid having to build a new data center. These benefits were in addition to the IT benefits of a reduced quantity of hardware to purchase, license, and maintain; savings in network ports and cards; a reduction in systems administration labor; easier recovery in the event of a disaster; and increased speed and responsiveness to changes in user capacity demand. When the virtualization project was first justified, the focus was on the IT benefits, not on the facility benefit of avoiding construction of a new data center.

Virtualization can refer to a variety of computing concepts, but it usually means running multiple operating systems on a single machine. While traditional computers have only one operating system installed, virtualization software allows a computer to run several operating systems at the same time.

the actual adoption rate of production servers, as opposed to an operator's being "invested in the technology," remains unclear.[26,27,28,29]

Even where virtualized resource pools are deployed, utilization is typically below 40 percent due to a combination of factors. Growth in hardware performance and capacity continues to outpace the consolidation of workloads. Moreover, data center operators tend to take a cautious approach to consolidation ratios, mindful of highly variable and unpredictable workloads and concerned about potential service level agreement (SLA) violations because performance could be affected by resource contention as server utilization increases with virtualization.[30]

To be clear, a utilization metric is difficult to increase because it gives as much weight to single-workload servers as it does to virtual hosts. So even with 90 percent of workloads virtualized, the remaining 10 percent could still account for half of all physical servers, thus having a large impact on the average utilization metric.

Another limitation to increasing utilization lies in the fact that not all servers, and not all the applications they run, can be completely virtualized. There are a number of reasons: A server may require physical peripheral hardware connections; a software license may not allow virtualization; time synchronization may be critical; or, where different departments own hardware or software, sharing of resources may pose a financial or political challenge.[31,32]

Reaching high utilization levels for a specific set of workloads, such as 50 percent or higher, may also entail high setup costs, which can outweigh the electricity savings benefit of higher utilization.

However, when applications and servers are designed with virtualization in mind, the potential for efficiency and cost savings is enormous. This is best illustrated by the hyper-scale cloud providers, which are able to treat their data centers as modern-day factories, optimizing efficiency wherever possible, because they often both own and operate the equipment and the entire data center. Consequently, they are highly incentivized to maximize productivity.

In summary, implementing virtualization well requires an investment of time and resources to properly design applications and storage. Incorrectly done, it can negatively affect IT availability and performance. Given this risk and other priorities, many IT professionals choose not to give virtualization the effort it requires, and end up doing nothing.

2.2.3 Unused ("comatose") servers

An estimated 20 to 30 percent of servers in large data centers today are idle, obsolete, or unused but are still plugged in and operating in "on" mode, consuming energy doing nothing. IT managers often find that they cannot identify owners for 15 to 30 percent of the installed server base but are reluctant to decommission equipment, fearing potential impacts on business or application functions. Many organizations also do not budget staff time to identify, remove, or repurpose obsolete servers.

A "comatose" (or "zombie") server is a server that is powered on and using electricity while delivering no useful information services. It may have been left behind after a project ended or after a business process changed and no longer required the application, but it has not been decommissioned, nor is anyone tracking its usage.

Numerous studies and surveys have determined that an estimated 20 to 30 percent of all servers are idle, obsolete, or unused but still consuming energy in data centers today.[33] The problem is likely more widespread than reported. According to Uptime 2013 survey data, only 14 percent of respondents believe their server populations include 10 percent or more comatose machines, yet nearly half of the respondents have no scheduled auditing for identifying and removing unused machines.[34] Common reasons for high rates of comatose servers include:

- lack of focus;
- aversion to risk; and
- split incentives.

Downsizing to reduce labor costs can result in poor maintenance and control of asset records. Some organizations do not budget staff time to identify, remove, or repurpose obsolete servers because they assume servers are inexpensive, often forgetting the electricity and facility costs associated with leaving them running. While software solutions are available, IT managers have been slow to adopt them or use them effectively to track idle servers. Even with software, identifying specific servers and the applications that run on them is a complex task.

Further, an aversion to risk makes IT managers reluctant to decommission any previously installed servers, fearing it may have some impact on a business or application function

Through industry initiatives like the Uptime Institute's Server Roundup, launched in 2011, there are now sponsored programs focused on raising awareness about the removal and recycling of comatose servers. However, the problem remains widespread across all data center types. In the 2012 Roundup, participants identified nearly 20,000 servers that were comatose, and shutting these off saved about 5 megawatts of IT load along with 4 megawatts of associated cooling and infrastructure load. AOL, winner of the 2013 event, decommissioned more than 9,484 servers for a total savings of close to $5 million.

that might still be running occasionally on the server. IT managers can lose their jobs for interfering with a business function, but to our knowledge no IT manager has been fired for keeping comatose servers online. Service level agreements may also make IT staff hesitant to turn off servers, for fear of affecting performance or availability.[35]

Organizations often do not realize the volume of idle equipment in their facilities until a data center must be closed and server processing moved to a new location, which is the only time application owners must be identified for all equipment. The remainder of the time, organizations consistently find that they cannot identify owners for 15 to 30 percent of the installed server base.

Removing comatose equipment is proving to be more of a management and behavior challenge than a purely technical one. While not immune, many hyper-scale cloud providers have a lower rate of idle servers because they have fixed the internal institutional problems that lead to separate budgets for the IT and facilities departments (split incentives) and dispersed responsibility for data center design, construction, and operations. These problems have not been corrected, however, for the vast majority of small- to medium-sized organizations with server closets, corporate enterprise data centers, and multi-tenant facilities.[36]

If IT departments paid the electricity and infrastructure costs for comatose servers, they would be more likely to ensure that these servers were decommissioned.[37] Unfortunately, Uptime's 2013 survey indicates that, for the third year in a row, the data center utility bill is paid for by facilities nearly 80 percent of the time, which makes progress on IT efficiency a serious challenge.

2.2.4 Inherent limitations to high utilization levels outside of public cloud computing

Multi-tenancy and dynamic provisioning provide public cloud computing with an inherent advantage when it comes to utilization levels.

The public cloud computing model has two major advantages over traditional data centers in terms of hardware efficiency: the ability to share IT resources to deliver a service for multiple customers—referred to as multi-tenancy—and dynamic provisioning, the scaling up and down of IT resources on an instantaneous basis.

Within small- to medium-sized organizations, enterprises, and often multi-tenant environments, IT managers must provision infrastructure against a forecast peak with a set number of servers, whereas cloud computing providers are able to match discrete peak demands against average demand by sharing infrastructure resources across its entire customer base. By aggregating the weakly correlated workloads of large numbers of customers and leveraging the concept of diversity, cloud service providers are consistently able to realize dramatically smaller peak-to-average ratios. This greatly reduces the need for idle reserve capacity and increases server utilization to rates well above 40 percent.[38]

Salesforce.com's cloud platform, as an example, combines the aggregation of workloads with a multi-tenant architecture that lets tens of thousands of organizations share what is logically a single infrastructure, running a shared version of software.[39] This integrated approach results in server productivity improvement (i.e., fewer watts to process the same information) that is instantly scaled across the entire infrastructure, achieving a significant multiplier effect in efficiency that reduces energy consumption and increases utilization.[40]

Traditional data center operators can and should increase their average server utilization levels to reduce energy consumption. However, moving their workloads to energy efficient cloud service providers may potentially be the most effective way to reduce their energy and environmental impacts.[41]

2.2.5 Under-deployment of power management software

Powering down unused servers could go a long way toward resolving the problem of over-provisioning. However, server power management solutions are still not widely adopted.

Current software offers cost-effective and simple means to measure, monitor, and manage hardware and application-level performance. Power management is a broad topic, and the industry separates it into data center infrastructure management (DCIM) and software to manage the IT hardware side. While DCIM is seeing increased adoption to improve the efficiency of cooling equipment, software solutions to manage IT hardware use are not as widely used. In part, this seems due to concerns by data center operators that DCIM technology may threaten their employment, leading some operators to discourage its adoption.[42]

Traditional DCIM focuses primarily on the measurement of power usage at the facility level and does not take application-level or IT workload into consideration. Additionally, it does not allow the actual automated control of servers. In comparison, power management for IT hardware can include the ability to put servers to sleep, as well as to monitor and spin up servers instantly from a sleep mode.

Companies such as TSO Logic, a relatively new entrant in the power management space, offer data center operators the ability to monitor server activity and utilization via software, down to the application level. Its software solution monitors power demand, cost of energy, and how these factors change over time, helping reallocate workloads and power down equipment.

Power-saving features embedded in the server hardware itself enable reduced power consumption in idle mode. These features also monitor and track hardware utilization, providing feedback to a data center dashboard that organizes, reports, and manages hardware functionality across the data center or at the customer rack and server level.

However, many IT managers disable these features when installing their own custom software images on the servers.[43] Even some hyper-scale cloud companies do not deploy full-scale power management due to the perceived complexity on the management side and risk aversion of powering down servers. Broad adoption across the market remains a challenge due to split incentives, which block the cost-saving benefits, or due to the fear of an impact on reliability.

There are also some technical limitations to reaching 50 percent average utilization on non-resilient servers with varied workloads, primarily due to memory resource issues under heavy virtualization. For most commercial servers, a realistic goal for average server utilization would be in the 20 to 30 percent range in the short-term, reaching toward 50 percent in the longer term (5 to 8 years). However even with these short-term technical limitations, there is still a large gap, and savings opportunity, between current average and best-practice utilization.

2.2.6 Procurement practices focus on first cost instead of total cost of ownership

When focusing exclusively on initial costs instead of lifetime electricity costs, IT procurement practices miss out on the most efficient servers available on the market.

IT manufacturers are progressively making server power consumption more proportional to CPU utilization. Power data from SPECPower show that servers have considerably reduced their power use at idle and low loads over the past five years.[44] This dramatically improves efficiency at common data center workloads, regardless of the utilization rate.

Intel's latest Xeon E5-2600 processor, which achieves lower power usage at low utilization, presents a compelling example. With an average power demand decrease of 50 watts since 2006, a company buying 10,000 servers annually over four years would see almost $10 million in electric utility OpEx savings over their life cycles, including facility total cost of ownership in the product selection process. More important, $56 million in data center construction could be deferred because of the servers' ability to handle increased loads at a lower power demand, thus realizing significant carbon savings in addition to cost savings.[45] However, modern servers are still far from power proportionality, and server efficiency gains yield limited energy benefits if their average utilization remains very low.

In addition, server efficiency is often not a top purchasing criterion. IT procurement officers typically look primarily at performance and cost first when selecting equipment. This makes IT energy efficiency a management and cost-accounting problem, not solely a technical one.[46]

2.2.7 Lack of a common, standardized server utilization metric

Until better metrics are available, CPU (central processing unit) utilization alone can be an adequate proxy to provide visibility on the IT efficiency of data centers, thereby creating market incentives for operators to optimize the utilization of their IT assets.

Average server utilization is far from a perfect proxy for IT hardware efficiency, but it does offer a cost-effective method available today to measure and evaluate basic IT performance.

And yet, average server utilization is not widely used as a key performance indicator due to the perceived complexity of collecting the data, a lack of awareness of its role in driving energy performance, or a belief that it is not sufficiently accurate. To be fair, CPU utilization is not a universal benchmark. Different workloads have different CPU intensities, with some being more disk-, network-, or memory-constrained than CPU-constrained. Most recent efforts to measure and improve efficiency have been on the facility infrastructure side. While having many beneficial effects, this does nothing to capture savings opportunities from IT efficiency.

Consequently, numerous metrics focused on IT efficiency have sprung up in recent years. For instance, Green Grid member working groups are developing a more sophisticated framing of a server utilization metric based on a handful of parameters that impact utilization: data center design, hardware, software, CPU versus memory versus network, and so on. Further, the Green Grid began another approach in 2011 by measuring efficiency at the application level, as many applications have counters to track how much work they are doing—tracking, for example, the number of emails processed or the number of users supported.[47] Yet another metric is focused on measuring the useful work out of the data center and how it relates to resource consumption.[48] While these metrics will increase the understanding of server utilization as it relates to workload and operational conditions, they do not provide a near-term, simple benchmark that can drive the optimization of server utilization across the data center industry as PUE does with facility efficiency.

Meanwhile, where data are available, we continue to see data centers report very low levels of CPU utilization, one of the primary determinants of IT efficiency and performance. As more complex standards continue to be developed and tested, CPU utilization alone can be an adequate proxy to provide visibility on the IT efficiency of data centers, thereby creating market incentives for operators to optimize the utilization of their IT assets and reduce energy consumption.

2.3 SPLIT INCENTIVES AND ORGANIZATIONAL MISALIGNMENT CONTINUE TO BE THE BIGGEST BARRIERS

Twenty percent of organizations' IT departments pay the data center power bill, a statistic that has not changed in more than five years. This misalignment of incentives is one of the major barriers to implementing efficiency measures in data centers. While hyper-scale cloud service providers have largely solved this issue, contractual relationships between multi-tenant providers and their customers compound the split incentive challenge, with the data center owner paying the power bill, the tenants buying power blocks, and their IT purchasers separately specifying equipment. Consequently, the disparate parties have little motivation to invest in efficiency.

Historically, many companies have placed IT and facilities in different parts of the organization, creating an inherent division of accountability and incentives for delivering data center energy efficiency. The IT department is typically not evaluated on the basis of how much energy its servers consume and may not make server efficiency a high priority, even though avoided infrastructure savings represent more than half of the economic savings associated with reducing IT electricity use in the data center.[49] When departments operate on separate budgets, they often do not feel the impacts of their decisions and are consequently not rewarded for any actions that drive efficiency.[50]

While this challenge has long been recognized, progress toward aligning IT and facilities has been very limited. According to the 2013 Uptime Institute survey, only 20 percent of organizations' IT departments pay the data center power bill, a statistic that hasn't changed in more than five years.[51] Ultimately, the problem is that the people who run data centers have little influence on fixing these institutional issues. Rather, it is the C-level stakeholders in the corporation (CEO, COO, CFO, CIO, etc.) who need to make these changes happen, and thus far there has been little movement in most companies. Once these problems are fixed, big changes in efficiency can follow and continue apace as they become part of the business culture and drive continuous improvements.[52] With proper alignment and incentives, IT managers can be rewarded for driving increases in server utilization and other IT efficiency improvements, in the same way that facilities managers are recognized for achieving cooling infrastructure improvements.

While the problem of split incentives remains a big issue for most in-house enterprise and small- and medium-sized organizations, it has been largely solved by the hyper-scale cloud providers, who generally have one data center budget and clear responsibilities assigned to an entity with decision-making authority.[53] Data center energy costs represent a large share of these businesses' operating expenses, making it a priority for them to optimize, whereas these costs are typically a small percentage of other companies' operational budgets and, as a result, fail to get the attention they deserve.

In 2012 eBay and the Green Grid released a report titled *Breaking New Ground on Data Center Efficiency*, which summarized how eBay's aligned IT and facilities organization positioned it to use a metrics-based approach to drive its data center design and server procurement process. The results were impressive, with a reported average 1.35 PUE, a best-case average PUE as low as 1.26 at 30 to 35 percent load, and hourly readings as low as 1.018 for stand-alone rooftop container data centers.[54] These results, however, were dependent on eBay aligning its organizational culture and accountability for data center capital and operational spending. With budgetary responsibility under one roof, incentives were aligned so decisions could be based on the total power consumption of data center buildings and hardware (i.e., the total cost of ownership, or TCO). In order to achieve higher efficiency in data center infrastructure and operations, eBay undertook two parallel but interlocking processes: technical and organizational optimization. Technical optimization required the simultaneous optimization of data center facilities, hardware, and software. At the same time, changes in eBay's IT organizational structure were implemented to drive cultural change with respect to building a more efficient infrastructure system, in order to take full advantage of newly implemented technologies.[55] An additional output from this effort was eBay's widely publicized Digital Service Efficiency dashboard.

In comparison, the multi-tenant data center contractual relationship between provider and customer creates an extreme case of the split incentive challenge, because the people paying the power bill and the IT purchasers work for different companies. Consequently, there is often little motivation to invest in more efficient equipment, let alone operate that equipment more efficiently.[56] In general, customers' IT departments specify the equipment needed to run their applications, while multi-tenant facilities staff are responsible for integrating those resources into the data center and providing reliable power and cooling. Because customers do not pay for the power or cooling required to operate their equipment directly, or because they are paying for large power blocks, they have little incentive to optimize the utilization of servers or invest in efficient equipment because it does not directly impact their cost of services.

2.4 MULTI-TENANT DATA CENTERS FACE ADDITIONAL CHALLENGES HINDERING ENERGY EFFICIENCY IMPROVEMENTS

Multi-tenant data center providers are struggling to implement energy efficiency best practices because of competing priorities, exacerbated split incentives, and challenges with incentive programs.

Our ever-expanding and changing digital society not only has spurred the growth and proliferation of data centers but has also caused the evolution of the data center and the industry itself. It is projected that more than a quarter of all data center capacity will be delivered by multi-tenant data center service providers by 2016, rather than being managed in-house by small, medium, and large enterprise organizations.[57]

Beyond the well-known hyper-scale cloud providers, multi-tenant data centers offer facilities and services that are well-suited to providing enterprises, small and medium-size organizations (SMOs), and other consumers with a bridge between fully owned and operated servers and data centers and fully outsourced cloud computing.[58] Many enterprises today are seeking out this model as their own data centers become obsolete or radically underpowered, or as they develop a need for non-capital-intensive expansion facilities and/or lower-cost backup data centers for disaster-recovery and business-continuity purposes.[59] As companies decommission or consolidate their existing data center space and instead procure this new type of service from third parties, the energy consumption implications are worth examining, especially since this market has not received as much scrutiny on its energy performance as have the hyper-scale cloud providers.

In addition to the challenges identified earlier, which apply nearly universally to data centers, there are three primary reasons why multi-tenant data centers have not been able to achieve best-practice energy efficiency performance to date, and they include behavioral, contractual, and policy-oriented barriers. Specifically, they are:

1. competing priorities due to the contractual relationship;
2. an amplified version of the split incentive between customers and provider, and related management/behavioral challenges; and
3. challenges in implementing efficiency incentive programs that are specific to multi-tenant data centers.

2.4.1 Priorities competing with energy efficiency

Multi-tenant data center providers have many different priorities in meeting the needs of their customers, keeping costs low, and maintaining high levels of security, reliability, and uptime. Focusing primarily on competing priorities negatively impacts energy efficiency.

In a rapidly growing, complex market, providers place a very high priority on reliability, security, availability, and customer service, and this tends to undermine interest in energy efficiency.[60] Multi-tenant data center providers make a business out of providing highly reliable infrastructure in which companies house the IT equipment they use for mission-critical applications. The redundancy and reliability of the data center, or the “tier” level, is an important metric for potential customers. In fact, many multi-tenant data center providers are selling higher levels (tiers) of reliability than their customers require, something that eBay identified as an unnecessary cost in its Project Mercury, during which the company moved 80 percent of its servers to a lower-tier facility.[61]

Security is another very high priority. Multi-tenant data centers require a high level of security because some of the information they contain is mission-critical, and customers tend to be more sensitive to data security when their information is outsourced than when they control it in-house. According to a major manufacturer of data center infrastructure equipment, these types of security concerns are sometimes used as an excuse by some multi-tenant providers to not implement simple savings solutions like hot/cold aisles and blanking panels, but the manufacturer claims these excuses are not valid and they are pushing multi-tenant providers to address this issue within their sphere of influence as added value for their customers.[62]

Multi-tenant customers are mainly concerned with the availability of their software applications. As a result, multi-tenant providers are extremely sensitive to anything that might affect the continuous operation of the facility, a challenge that is compounded once a facility is operational and populated with customers. Many interview respondents reported their perception that energy efficiency was a low priority for customers; some indicated that customers never even asked about the efficiency of the facilities or inquired about energy efficiency upgrades. It is thus unsurprising that providers are not motivated to pursue energy efficiency over other business interests. One interviewee, a multi-tenant hosting provider of cloud services operating in a wholesale data center, didn’t know the PUE of his facility, nor did he seem interested in determining it.

2.4.2 Amplification of split incentives in the multi-tenant data center environment

Prevailing space and power-block pricing models offer little incentive for customers to purchase more-efficient IT hardware, and few providers offer transparent energy reporting to customers. However, a small but growing number of providers report and charge for actual space and energy use, which can help align incentives and improve efficiency.

The contractual relationship between multi-tenant data center providers and their customers creates an extreme case of the split incentive because the people paying the power bill and the IT purchasers work for different companies. Without pricing contracts that reflect the costs of powering and cooling the equipment, there is little motivation for multi-tenant data center customers to invest in more-efficient equipment or operations.[63]

A retail provider of multi-tenant services commented that "the barriers to a per-kWh pricing model are the perception that it will cost extra for the provider to monitor equipment and [the knowledge] that the contract would need to be changed and terms adjusted—the latter being a meaningful barrier in terms of time and paperwork. However, if we were able to, through consolidation, recoup the space and power and allocate that to other customers, the economics would make it beneficial for both parties."

Multi-tenant data center pricing models directly affect the ability to save energy from the perspective of both the service provider and the customer. The pricing models defined below differ primarily in how they address space (physical square footage or rack space), power, cooling, and other services.[64]

a. **Space-Based Pricing:** Customers pay by the rack or by square footage and are usually limited to a certain power budget that is included in the space charges, with additional fees if they use power over that limit. With this pricing model there is little financial incentive for the customer to lower energy consumption unless approaching the power limit for the given space. Space-based pricing used to be the standard pricing model for the industry. However, as data centers' ability to grow becomes increasingly limited by the amount of power and cooling available (rather than by physical space constraints), this pricing model is being replaced by different models that more directly factor in energy use.

b. **Space and Power Block Pricing:** Some multi-tenant data center providers use a combination of power-based pricing and space-based pricing. In these models, the customer pays for a certain amount of space, and pays for power separately. These power charges are typically based on a maximum allowable draw. When power is charged separately but by the block up to a maximum wattage, customers have less of a financial incentive to reduce their energy use, unless they are near the wattage limit beyond which they will incur higher charges.

c. **Space and Actual Power Pricing:** A small number of multi-tenant providers use a combination of power-based and space-based pricing, but rather than charge for blocks of power, they charge on the basis of actual energy consumption, which can vary. In this case, customers have a direct financial incentive to reduce their energy use, as it directly lowers their operating costs. This can be achieved by increasing the efficiency of the hardware and also by consolidation, which reduces both space and power charges. The provider also can realize a financial benefit, either by requiring less cooling (and therefore less energy demand) for the same customer or by maximizing space use, allocating the freed-up space to another customer.

d. **Cost-Plus Pricing:** A small but growing number of facilities and providers document all costs to operate the data center and then charge customers on the basis of these charges, with an automatic markup.

As power densities continue to increase, data centers are becoming increasingly power-limited. Consequently, a few facility providers are switching to pricing models under which customers pay more directly for the energy they consume. While the standard contract does not promote efficiency today, nearly one-third of all leased data center space in the United States will come up for renewal over the next year, with nearly all of those leases being based on square footage.[65] If data center providers and their customers are going to recognize together the cost savings of better efficiency, the time to do so is now.

An issue related to the various pricing models offered by multi-tenant data center providers is the widespread lack of reporting, which is a result of customers not seeking additional information beyond cost, and providers not disclosing energy performance metrics that could impact how customers configure, operate, or optimize their IT hardware and procured services. This creates an information asymmetry that has allowed providers who use block pricing to charge for power that is going unused. For companies that have consolidated their own data centers and moved to the cloud, a simple bill is often a relief when compared with the complexity of managing a data center. In this transition, however, many companies are losing the granularity that provided insight into their systems' performance. Consequently, unnecessary energy consumption may be escalating out of control, or at the very least going unmeasured. With advances in DCIM and power management software, providers are now able to supply analytics at a very detailed level at a cost that is steadily dropping. But unless customers demand increased reporting around the energy, carbon impact, and performance of their purchased services, providers are unlikely to disclose this information.

2.4.3 Challenges for utility incentive programs and financing specific to multi-tenant data centers

The business relationship between multi-tenant data center service providers and customers makes it challenging for utilities to offer effective energy efficiency incentive programs. Additional challenges include the need for shorter-term ROI (return on investment) and the complexity in measurement and verification of implemented energy savings projects.

While many utilities and local governments have offered tax breaks and other incentives to bring hyper-scale cloud providers to their regions, few utilities manage energy efficiency incentive schemes to promote operational efficiency.[66,67] Utilities' ability to engage data center providers on energy efficiency practices is often limited in multi-tenant environments because they are restricted to offering incentives to customers of record. This is the case, for example, with Silicon Valley Power (SVP), which offers its direct customers energy efficiency incentives and is allowed to recoup the incentive amounts if a measure does not stay in place for the contracted five-year period. However, SVP cannot give incentives to a multi-tenant service provider's customers, as these customers do not pay SVP for their power directly. This is a major barrier for many utilities attempting to persuade multi-tenant service providers to participate in virtualization incentives.[68] Despite this barrier, some utilities have been successful, such as Consolidated Edison in New York State, in collaboration with the state's Energy Research and Development Authority (NYSERDA); and Seattle City Light, which has supported a number of multi-tenant service providers in Washington State to consolidate space, virtualize servers, and invest in improved cooling technologies.[69]

Some multi-tenant service providers in our survey mentioned a number of financial barriers to energy efficiency upgrades. Several interviewees indicated the ROI of efficiency projects is often not fast enough, as companies tend to look for an ROI of less than two years to justify a project, especially an upgrade. This seems especially to be a problem in regions with cheap power, which results in a longer ROI for efficiency upgrades. While efficiency measures can be easily implemented at the data center infrastructure level, aided by measurement and verification standards to isolate improvements and track performance (and potential incentives offered by utilities), there remain both technical and management barriers to implementing efficiency projects at the hardware and software levels.

3. RECOMMENDATIONS

The move to the cloud is clearly poised to continue, driven by economies of scale and growing acceptance of the security, accessibility, and improved cost efficiency of this business model. As the factories of the digital age, data centers require substantial amounts of energy while also presenting major opportunities for dematerialization and decoupling of economic growth from environmental impact. While the industry as a whole continues to make progress with many of the challenges identified in this report, there remain a number of strategic and tactical barriers that need to be prioritized to curb data center energy growth. Further, business-as-usual incentives and operating practices are proving insufficient to drive efficiency in the multi-tenant data center market, which has been largely ignored thus far by stakeholders concerned with energy consumption in the technology sector at large. To move forward, a suite of prioritized, strategic actions are necessary to accelerate the known mechanisms of efficiency and enable faster adoption of efficiency-focused technologies, metrics, and initiatives in all data centers.

If many of the recommendations presented here are successfully adopted, the direct energy, carbon, and cost savings would be significant. To illustrate, we turn to research published in 2011 by Masanet et al. estimating that the technical electricity savings potential in U.S. data centers in 2008 was 80 percent.[70] The measures analyzed in Masanet's research are largely similar to the ones suggested in this report. While we don't know how much savings potential would result from removing each of the market barriers discussed in this report, if enough barriers were removed to realize half of the technical savings potential for data center efficiency, electricity consumption for U.S. data centers could be cut by as much as 40 percent. In 2014, this represents a savings of 46 billion kilowatt-hours annually, equivalent to the annual electricity consumption of nearly all the households in the state of Illinois. Such improvement would save U.S. businesses $4.5 billion annually.[71] These results on their own suggest both widespread inefficiencies in current data center operations and the availability of technologies and operating practices that could reduce these inefficiencies significantly. While IT efficiency accounts for 45 percent of the total savings, its compounding effect on the facilities side can be seen through the reduced demand in infrastructure equipment electricity use: Nearly 80 percent of the infrastructure electric savings would result from reduced IT demand.

3.1 ADOPT A SIMPLIFIED CPU UTILIZATION METRIC TO ADDRESS THE BIGGEST EFFICIENCY ISSUE IN DATA CENTERS: UNDERUTILIZATION OF IT ASSETS

A number of metrics are available today that address aspects of IT equipment efficiency, but adoption is slow because they are often too complicated to implement consistently, easily, and cost-effectively and to report to management. Notwithstanding certain limitations, measuring and reporting CPU utilization is recommended as an adequate, simple, and inexpensive-to-monitor proxy of IT efficiency that could be used in the short term to promote greater disclosure and transparency and ultimately to drive greater IT energy efficiency in data centers.

The practice of measuring data center efficiency got off to a promising start when the Green Grid came out with the power usage effectiveness (PUE) metric in 2007. PUE has become the touchstone for measuring data center energy efficiency because it was the industry's first metric that was easy to understand and apply. While many data centers now report PUE and are making progress to improve it, the industry quickly recognized that PUE only scratched the surface in evaluating total data center efficiency. Numerous alternative metrics have sprung up in recent years to address the IT hardware side of data center energy use. Among the more widely recognized are these:

- **Corporate Average Data Center Efficiency (CADE):** This is a set of four metrics, developed by the Uptime Institute and McKinsey & Company in 2008, that measure facility asset utilization, facility energy efficiency, IT asset utilization, and IT energy efficiency. While innovative at the time, CADE has not been widely adopted, likely due to its complexity.
- **Power to Performance Effectiveness (PPE):** Developed by the market research firm Gartner, PPE measures server performance per kilowatt with the goal of helping IT managers raise server utilization levels.
- **PAR4:** From Underwriters Laboratories and Power Assure, PAR4 calculates transactions per second per watt.
- **SPUE (Server PUE):** Proposed in the lecture series The Datacenter as a Computer, SPUE is the ratio of total server input power to its useful power, where useful power includes only the power consumed by the electronic components directly involved in computation: motherboard, disks, CPUs, DRAM, I/O cards, and so on.

- **SPECpower_ssj2008v1.12:** The Standard Performance Evaluation Corporation (SPEC) developed this metric, the first industry-standard benchmark that evaluates the power and performance characteristics of single servers and multi-node servers. While not a metric on its own, this benchmark provides a means to measure power (at the AC input) in conjunction with a performance metric, and it can help IT managers consider power characteristics, along with other selection criteria, to increase the efficiency of data centers.
- **Data Center Productivity (DCeP), Compute Efficiency (DCcE), and its sub-metric, Server Compute Efficiency (ScE):** Proposed by the Green Grid, these would enable data center operators to determine the efficiency and productivity of their compute resources. DCcE is not a productivity metric on its own, as it does not show how much useful work is being done by the data center, but it can be used in conjunction with other proposed DCeP metrics developed by the Green Grid to evaluate the proportion of measured work (e.g., number of emails delivered, GB of data stored, transactions processed) that the data center is delivering compared with the amount of energy consumed.[72,73] However, DCcE is not a one-size-fits-all metric, and it should not be used to make comparisons between different data centers delivering different types of services. While the Green Grid is beginning to promote performance-based metrics, a straightforward efficiency metric remains elusive.

Other metrics focus on performance per watt or carbon emissions per transaction (or per user, per email, and so on). All of these metrics address aspects of IT equipment efficiency but are often too complicated for IT managers and data center operators to implement consistently, easily, and cost-effectively and to report internally and externally. Following its work on PUE and other productivity-based metrics, the Green Grid is currently in the process of establishing additional metrics focused on utilization of specific aspects of the data center (e.g., compute, network, memory, applications), but these have yet to be widely adopted. Overall, while there is general acknowledgment of the need to apply more consistent and meaningful metrics to inform data center efficiency, it remains a challenge, and there continues to be a gap in the consistent use of metrics beyond PUE.

3.1.1 Toward a CPU utilization metric

We recommend a standardized, simple, and inexpensive approach to monitoring CPU utilization, which would support greater disclosure, transparency, and reporting by companies with data centers.

While CPU utilization alone does not determine the usefulness of the work output, it would be an adequate proxy for the time being, while more complex and interrelated metrics are developed. A former executive of the Green Grid believes “the issue of utilization has been overcomplicated. Utilization as a metric itself is extremely useful, and it’s not that hard to measure on a broad basis.” He added that “focusing on CPU utilization of a given population of servers is a good way to understand overall system performance, as it’s highly correlated to end-user productivity. And since every machine can report CPU utilization within a specific type of application or usage, you could compare like for like performance.” Thus, CPU utilization can support greater disclosure, transparency, and reporting by companies with data centers and support many of the complementary recommendations suggested here.

The following two metrics can best help to track and report server utilization in data centers:

Average Server Utilization: average CPU utilization of a server over the period of time considered (day, month, year)

Average Data Center Utilization: unweighted average of Average Server Utilization across all servers (not in sleep or off mode) in owned and operated data centers

The broad adoption of these simple utilization metrics across the data center industry would provide visibility on the IT efficiency of data centers, thereby creating market incentives for operators to optimize the utilization of their IT assets. It could spur dramatic increases in data center IT efficiency, just as PUE did for facility efficiency.

This proposed data center utilization metric could be expanded to include metrics for the utilization of storage and networking equipment, as well as more sophisticated methods that are currently in development to measure useful work and data center productivity. However, CPU utilization remains the largest IT electrical load in data centers and should therefore be the first priority.

An immediate benefit of a server utilization metric would be to facilitate “server roundups” to identify, assess, and decommission comatose servers, as advocated by the Uptime Institute. Uptime’s Server Roundup contest has shown that decommissioning a single one-rack unit server (1U) can result in a savings of $500 per year in energy costs, an additional $500 in operating system licenses, and $1,500 in hardware maintenance costs. Since the contest’s launch two years ago, participants have decommissioned and recycled 30,000 units of obsolete IT equipment and saved millions of dollars in annual operating costs.

Multi-tenant data center providers and customers have also proved they can successfully engage with utility and energy efficiency programs that offer incentives for improvements in hardware efficiency (along with infrastructure cooling efficiency efforts that are commonplace today, such as those offered by NYSERDA). At Seattle City Light, an incentive program enabled one of the utility's customers to upgrade its uninterruptible power supply (UPS), and while preparing for the transition the customer identified 13 percent of its servers as comatose or abandoned.

3.2 INCREASE DISCLOSURE OF DATA CENTER ENERGY AND CARBON PERFORMANCE

As DCIM tools provide more insight into data center performance, IT managers should use the data to benchmark and report performance to internal and external customers and other stakeholders. This is a critical step in raising awareness and engaging stakeholders in finding solutions to improve energy efficiency.

Adopting DCIM tools to manage and report energy and carbon performance will dramatically increase the data center industry's ability to assess performance with a suite of key indicators. And with the alignment of IT and facilities departments, management will demand that these metrics be produced and acted on. Further, as data center footprints expand and energy consumption rises, organizations are increasingly under pressure to reduce the cost of their operations and to report and reduce their environmental impact. In many regions of the world, governments and corporations now require a measurement of the carbon footprint of services and products, in addition to operations. For example, the CDP (formerly the Carbon Disclosure Project) is persuading businesses to issue corporate responsibility reports that include carbon reporting and, at a sector level, voluntary reporting of data center energy consumption and performance. Correspondingly, releasing data center activity and performance data will support better decision making, competition, and adoption of best practices in the industry, leading to improved performance and a smaller environmental footprint. Further, the recently released draft reporting standards from the Sustainability Accounting Standards Board (SASB) for the information and communication technologies (ICT) sector recommends that IT companies with data centers report material metrics like weighted average PUE for owned and operated data centers, total energy consumption of data centers, and descriptions of current environmental criteria for determining the location of new data centers, including factors affecting energy and water consumption.[74]

Although data centers can be located anywhere on the globe, rarely do providers evaluate the carbon intensity of the power being delivered to their facilities, which has a significant impact on data centers' overall carbon emissions impact (though a focus on IT efficiency remains the single largest strategy to reduce total carbon emissions).[75] Rather, providers prioritize the availability of competitively priced power, the availability of high-bandwidth Internet backbone connectivity, and proximity to regional network hubs. Thus, in addition to the day-to-day performance metrics that DCIM enables, data center siting criteria and policies that consider environmental, as well as social, impacts can drive the industry toward a more sustainable outcome. Recent efforts by Greenpeace have raised the profile of this issue in particular, and there is now a Future of Internet Power initiative organized by the nonprofit business network Business for Social Responsibility, though there are currently no multi-tenant data center companies participating.

Once again, the hyper-scale cloud providers are taking a leadership position by disclosing their data center operational performance with metrics that extend beyond PUE, enabled by integrated management systems and tools. Driven either by strong performance or by stakeholder influence (such as through Greenpeace's Cool IT program), the rest of the industry should follow this lead and incorporate green IT and data center reporting in similar ways. Industry organizations like the Green Grid are driving the development and standardization of common information models to profile the key attributes associated with energy and carbon emission computation and reporting, such as power and CPU profiling, and are continuing to recommend methodologies (beyond PUE) for improved efficiency reporting. When discussing the need for simple metrics to evaluate efficiency, an Uptime Institute leader commented that "CPU utilization as a metric alone may be good place to start consistently reporting IT performance, but it's more important for companies to measure and report something consistently over time so they can interpret performance relative to the service. No one number can do it all, but some number is better than none." In addition to SMOs and corporate enterprises adopting a more transparent reporting mechanism, cloud and multi-tenant data center service providers could integrate these types of metrics into billing and monitoring processes to support a transparent and per-tenant view of carbon footprint emissions and other key performance indicators of efficiency.

Beyond the core set of traditional performance metrics that consider cost and related factors to evaluate data center performance, a suite of reportable metrics that inform data center operational energy and carbon efficiency could include the following:[76]

Table 2	
Metric	**Reference/Source**
Annual energy consumption and carbon emissions for owned and operated data centers and/or services (total for facility, and/or broken down by customer)	CDP ICT Sector Module, SASB
Power usage effectiveness (PUE), carbon usage effectiveness (CUE), and water usage effectiveness (WUE) ratios as weighted averages for owned and operated data centers	Green Grid, CDP ICT Sector Module, eBay DSE, WRI/WBCSD GHG Protocol ICT Sector Supplement
Average server utilization and average data center utilization for owned and operated data centers	This study
Weighted average carbon emissions factor for consumed electricity by owned and operated data centers (purchased and/or generated)	Greenpeace, US EPA/ IEA, WRI/WBCSD GHG Protocol ICT Sector Supplement
Useful work (e.g., transactions, URLs) per kWh	eBay DSE, WRI/WBCSD GHG Protocol ICT Sector Supplement
MW from low- or no-carbon emissions sources (by type)	eBay DSE
Report on data center siting criteria	SASB
Report on measures planned or completed to increase energy efficiency of owned and operated data center(s), including any targets or goals	CDP ICT Sector Module

3.3 ALIGN INCENTIVES BETWEEN DECISION MAKERS ON DATA CENTER EFFICIENCY

Integrated business models wherein IT and facilities teams share incentives and implement common charge-back mechanisms are foundational to scaling virtualization, eliminating comatose servers, and procuring more efficient hardware.

The corporate disconnect between IT and facilities operations continues to challenge the data center industry. To effectively roll out energy efficiency programs across the IT side, in particular, data center managers need to overcome this organizational barrier and get executive-level buy-in in order to implement effective server decommissioning programs and other activities. A single business model wherein IT and facilities departments work together and let the functional requirements and economics drive the solutions, and customers are charged for energy consumed, will be critical to harness the power and cooling resources within the data center. Only by establishing a tightly integrated group that combines IT and facilities can an organization fill in the gaps and improve the overall business process. Doing so will break down the management barriers, which pose a greater challenge today than do the technological barriers. Technical solutions like increasing server utilization rates, scaling virtualization, eliminating comatose severs, and procuring more efficient hardware are well known but have not been realized at scale due to the misalignment of accountability and incentives.

3.3.1 Align contract incentives between multi-tenant data center providers and customers

Addressing the contractual issues that exist around pricing for data center services by moving toward actual space and energy use charge-back mechanisms, along with improved reporting and greater transparency, will help align the incentives required to realize efficiency gains.

Among multi-tenant data center providers specifically, addressing the contractual issues that exist around pricing for cloud and data center services, and improving the information exchange between providers and customers, are key actions that will drive increased efficiency. Once there is money on the line for either or both parties, decisions can be made inclusive of energy performance. Different multi-tenant service providers have different pricing models and methods for allocating charges for power, cooling, and space, and this differentiation is clearly a competitive component. But as data centers become more constrained by power and cooling (and less constrained by space), some facilities have proved that moving away from charging for power blocks and instead charging directly for power and cooling on a per-kWh basis can create more of an incentive for their customers to save energy.[77] Interviewees have suggested that this creates positive differentiation in the market, and increasingly, customers are requesting a switch or moving to providers who can offer this model.

Finally, there are some states, such as Washington, that have legislation in place to mandate a per-kWh charge instead of block charging. Many of the larger, wholesale, multi-tenant data centers are now offering this pricing model, as well as offering the tenant ownership of, and responsibility for managing, part or all of the cooling and power infrastructure for the space it leases. This could include computer room air handlers and air conditioners (CRAH/CRAC), chillers, packaged rooftop units, power distribution units (PDUs), and UPS units. Some multi-tenant data centers also offer the customer the ability to tie into their chilled water plant or their central condenser loop for heat rejection, and in either of these cases, the tenant (a colocation provider or enterprise customer) often does have an incentive to improve efficiency. Examples like these are promising, if they can scale to other regions and throughout the data center industry.

3.3.2 Develop a green multi-tenant data center service contract template

Development of a green data center service contract raises awareness, incentivizes energy efficiency, and facilitates implementation.

By making customers responsible for the power used by their equipment, green data center service contracts create a financial incentive for tenants to consider energy use in their decision-making process. To support this effort, industry organizations like the Green Grid and others are considering the development of contractual templates that would serve as a blueprint for customers and providers to use in pricing contracts. Similar to the green lease in the real estate sector, a green data center service contract could dramatically scale up this effort and lead to a number of key benefits beyond the financial advantages of increased performance. These include improving the environmental performance of leased space by securing critical commitments from both the provider and the customer, and improving environmental data reporting transparency to enable customers and providers to measure achievements against agreed-upon goals and metrics.

3.3.3 Adopt and deploy DCIM and power management software to enable efficiency and reporting

A higher rate of adoption and widespread deployment of DCIM and power management software tools are needed across all data center types to enable transparency and effective and efficient reporting, as well as to monitor the deployment of IT and facilities energy efficiency best practices.

To support the alignment of IT and facilities business functions, or—in the case of multi-tenant data center providers—to enable new contracting models and manage customer installations in a more granular way, data centers must adopt and deploy DCIM and power management software at levels beyond what we see today. DCIM enables the visualization of the data center with the ability to manage resources to increase uptime on one side and improve airflow on the other, creating more computing cycles. DCIM fundamentally is about adding more interactivity to the physical world that traditional IT runs on. A great example of a cross-organizational deployment of DCIM is eBay's much publicized Digital Service Efficiency (DSE), which is a finely tuned monitoring fabric, coupled with the requisite business logic and presentation layers, to identify key performance indicators and other metrics that make sense to various stakeholders.

Application-specific DCIM provides a more comprehensive solution that is customized for cloud and multi-tenant facilities. It provides facilities, IT, and other executive officers with detailed, real-time analytics, risk awareness, data center capacity management, and operation status to initiate automated actions or to recommend manual actions. This helps optimize IT performance, identify and eliminate comatose servers, and decrease downtime. Application-specific DCIM for multi-tenant data centers also provides insight into usage and availability on the tenant level, enabling capabilities like detailed charge-back for power and integrated tenant billing and support.[78]

How a better understanding of its energy use led a cloud services company to save tens of millions of dollars

A global cloud services provider that uses multi-tenant data centers to host its entire infrastructure recently discovered, through an initiative to increase the accuracy of its carbon emission calculations and reporting, that it was using less than half of the energy it provisioned in flat-fee pricing contracts with its multi-tenant providers. This realization and a comprehensive evaluation of its own hardware requirements gave the company a clear incentive to transition its energy contracts from block to a metered, or use-based, pricing model. This contractual change has to date resulted in more than $20 million in energy cost savings annually.

> "We make design decisions based on assumptions when we don't have data. Once we have data, it's used to validate our assumptions," explained the company's sustainability officer. "The data can expose not only the flaws in our operational design, but also opportunities for better design and improved efficiencies, and help inform our understanding and strategy."

The change to use-based energy pricing incentivized the company to look not just at peak power, but at all energy--saving opportunities, including use of off-peak energy. This new focus led to a new appreciation for the value of real-time energy data collection. Scrutiny of real-time data inspired and unlocked additional innovation around operational efficiency. For example, the company found it was being overly conservative in limiting the number of servers in each rack to avoid perceived maximum power constraints. In addition, racks were becoming de-optimized from a power perspective as servers were moved around over time. The increased level of data collection and performance monitoring gave operations managers a detailed understanding of hardware operating parameters, allowing them to increase server density per rack and thereby reduce multi-tenant rental costs with no fundamental changes to their platform. This initiative realized several million dollars in savings in multi-tenant provider costs during a period of 40 percent growth in infrastructure.

Real-time data metering has enabled a better understanding of business operations and the potential for savings from managing server power. The identification and power management (putting to sleep) of underutilized servers represents operational cost savings of millions of dollars per year. Real-time metering has also facilitated the modeling of new server deployment projects that will enable contract managers to confidently provision for actual power requirements, avoiding costly overprovisioning. The company is currently working on these initiatives.

4. MAKING IT HAPPEN: RECOMMENDED ACTIONS

Federal and State Governments

1. Set government procurement policy around multi-tenant data center pricing models that require charging by space and energy use.
2. Set renewable portfolio minimum standards for large data centers.
3. Develop a green data center contract template for the multi-tenant data center segment that promotes basing charges on space and energy consumption.
4. Lead by example: Disclose data center utilization metrics and require that government contractors do the same.
5. Align utility business incentives with energy efficiency by decoupling utility revenues from electricity sales, allowing them to profit from energy savings instead.

Electric Utility Companies

1. Implement energy efficiency incentive programs for data centers, including programs focused on improving equipment utilization, and programs targeting multi-tenant data centers.

Hyper-scale Cloud Service Providers

1. Continue leadership role in advancing infrastructure and IT efficiency best practices.
2. Transition portfolio of data centers to procure and/or generate renewable energy.
3. Disclose data center utilization metrics.

Multi-tenant Data Center Providers

1. Adjust pricing models to incentivize energy efficiency by customers.
2. Adopt and deploy DCIM tools to increase energy performance and improve reporting.
3. Disclose data center energy performance metrics (such as those proposed in this report) and incorporate into billing information.

Multi-tenant Data Center Customers

1.In new contracts, negotiate for pricing models based on actual space and energy use, and explore options to renegotiate existing terms.
2. Develop and pilot a green data center contract template.
3. Request data center energy performance metrics, and consider energy and environmental performance as a part of the total cost of procuring data center services.

Enterprise and Small- to Medium-sized Organizations

1. Have the sustainability or environmental department engage with the IT and facilities teams to quantify the energy and environmental impacts of data centers.
2. Obtain c-suite (top managers) buy-in to align IT and facilities organizations to drive energy efficiency across the data center.
3. Disclose data center utilization metrics.
4. Consider modular data centers to improve operations and services.

Industry Associations and NGOs

1. Issue best-practice guidance, standards, and methodologies, and push higher-profile leaders to disclose progress—and make further advances—on energy efficiency.
2. Focus on engaging with and driving best practices in the fastest-growing segment of the data center industry, multi-tenant data center providers, and also in the largest segment of the industry, SMO and enterprise data centers.
3. Develop a green data center contract template for the multi-tenant data center segment.

All

Promote a three-tiered strategy to reduce energy use in data centers by focusing on IT efficiency, infrastructure efficiency, and low-carbon power sources.

APPENDIX 1

MULTI-TENANT DATA CENTER DELIVERY MODELS: DEFINITIONS

Within the multi-tenant data center industry, there are a number of delivery and service models that are important to differentiate.[79]

1. **Wholesale multi-tenant facilities:** Wholesale data center providers lease data center space that is typically sold in cells or pods (individual white-space rooms) ranging in size from 10,000 to 20,000 square feet. Wholesale data center providers sell to both enterprises and multi-tenant providers.[80] Customers then engineer private data centers within these facilities or offer other types of data center services. The best way to think of a wholesale data center is as "data center space for rent." It is similar to leased office or warehouse space where the landlord provides facility maintenance services.

   Commercial arrangements: In a power-based arrangement, wholesale customers invest in power and cooling equipment with near-absolute control over data center layout, infrastructure, and ongoing management. In a turnkey arrangement, the IT infrastructure is owned and managed by the customer, but the cooling and facility infrastructure is owned and managed by the wholesale facility provider.[81] By design, wholesale providers do not have access to, nor the ability to control, a customer's hardware environment, and often, customers may not have any information on infrastructure efficiency or data other than their own power draw.

   Segment leaders: Wholesale multi-tenant providers often operate as real estate investment trusts (REITs), with the largest companies in the sector including Digital Realty Trust, DuPont Fabros, CoreSite, and Quality Technology Services (QTS). Other recent entrants include ByteGrid Holdings, Carter/Validus, DataFoundry, DataBank, Lincoln Rackhouse, and Power Loft.[82]

2. **Retail multi-tenant (colocation) facilities:** Multi-tenant data centers can be thought of as "retail" data center space that is leased on the basis of individual racks/cabinets or cages. Cages typically range from 500 to 5,000 square feet. Multi-tenant providers sell to a wide range of customers, from Fortune 1000 enterprises to small- and medium-sized organizations.[83] Typically the facility provides power and cooling to the space, but the IT equipment is owned by the customer. The value proposition of retail multi-tenant is that customers can retain full control of the design and management of their servers and storage, but turn over the daily task of managing data center and facility infrastructure to their multi-tenant provider.[84]

   Commercial arrangements: Multi-tenant facilities typically use a combination of three main pricing structures: space-based pricing, power-based pricing, and cost-plus pricing. These pricing models directly affect the ability of the provider and the customer to save energy and also determine any incentives to do so. The pricing models differ primarily in how they address space (physical square footage or rack space), power, cooling, and other services.[85]

   Segment leaders: Equinix, Telecity, and Interxion are three of the leading companies that compete with subsidiaries of large telecommunications companies like AT&T, Verizon, and Level3. There are also numerous larger, private network–neutral multi-tenant providers (i.e., providers that allow interconnection between multiple telecommunication carriers from the data center to the user) that compete in the market, including i/o Data Centers, Latisys, Telehouse, and the Telx Group.[86]

3. **Hosting facilities and services:** There are three kinds of hosting facilities—managed hosting, dedicated hosting, and shared hosting. Managed services are the most complex and flexible of the three. In addition to providing a server (or a part thereof) and storage for a client, managed hosting adds significant administration and engineering services to the mix. In a managed hosting environment, either the hosting provider or the customer may own the server hardware. Managed hosting providers may own their data centers or lease space in a multi-tenant facility.

   Dedicated hosting involves the leasing or rental of a server by a customer. That server is "dedicated" to a single customer—it is not shared, regardless of how it is used, and therefore it is also often referred to as a private cloud. Customers exercise full control over the server, beyond physical maintenance, and the hosting provider generally does not include administrative services beyond ensuring the server's ability to function at the hardware level.

   In shared hosting, many customers share a single server and operate a variety of applications from that server. Many smaller and midsize shared hosting providers will purchase services from dedicated hosting providers and use the dedicated servers to handle a large number of shared accounts, without infrastructure ownership.[87,88]

   Commercial arrangements: Hosting services are typically offered on a lease basis, depending on the power draw or specific application services that are being provisioned. In dedicated hosting where the customer is directly leasing a standalone server, a space-based, power-based, or cost-plus pricing model may be applied.

Segment leaders: Managed hosting leaders in business units within larger network service providers include AT&T, Time Warner Cable/NaviSite, CenturyLink/Savvis, Cable & Wireless, Orange Business, and Verizon Business/Terramark. Other top competitors are divisions within global IT professional services firms like HP/EDS, IBM, and CSC. Major private competitors with positions in both managed hosting and cloud services include Layered Tech, Softlayer, and SunGard, the global leader in disaster recovery. Other top-managed hosting companies in North America include Peak 10, Hosted Solutions, Carpathia, PEER 1, Q9, Hosting.com, ViaWest, and Verio.[89]

4. **Hyper-scale cloud computing:** Cloud computing is a model of service delivery for individual servers and storage allotments. It has a broadly accepted definition first put forth by the National Institute of Standards and Technology (NIST): Compute (server instances running an operating system), storage, and application resources must be available over a network as a shared pool of resources; it must also be elastic (able to be grown and shrunk by the user automatically), have automatic and real-time billing, and be self-service—that is, available to an end user to provision. There are three canonical categories: infrastructure as a service (IaaS), which is compute and storage provisioned to a user via API or portal; software as a service (SaaS), an application delivered online from a provider; and platform as a service (PaaS), an environment designed to let developers interact with code and create running applications without maintaining or operating the run time.[90] Cloud computing providers can own and operate their own data centers or lease data center space through any of the multi-tenant models (wholesale, retail, or managed/shared hosting).

   Commercial arrangements: Cloud computing services are typically offered on a per-user, per-application, or per-seat license basis.

   Segment leaders: Cloud computing leaders include some of the most public-facing companies on the Internet today and provide both business and consumer services for applications like email, storage, photo sharing, music, customer relationship management, and collaboration services. These include Microsoft, Apple, Amazon, Facebook, Salesforce.com, and many others.

APPENDIX 2

U.S. Data Center Segmentation Energy Use Methodology and Assumptions							
Segment	**% of stock (based on # of servers)**	**Average PUE**	**Average server utilization**	**Average server age (years)**	**2011 Electricity Use (MWh)**	**Server power at average utilization level (SPECpower_ ssj2008) (watts)**	**DC market segmentation by electricity consumption**
Small- to Medium-sized Data Centers	40%	2.0	10%	3	37,500,000	149	49%
Enterprise/ Corporate	30%	1.8	20%	2	20,500,000	120	27%
Multi-tenant Data Centers	22%	1.8	15%	2	14,100,000	113	19%
Hyper-scale Cloud Computing	7%	1.5	40%	1	3,300,000	101	4%
High-performance Computing	1%	1.8	50%	2	1,000,000	169	1%
	100%				**76,400,000**		**100%**

Notes:

We are not aware of existing estimates of U.S. data center energy consumption per the market segmentation used in this study. This model is meant as a rough estimate to help readers understand the relative contribution of each segment to the energy consumption of data centers in the United States. It is based on available data, wherever possible, and on the authors' assumptions where we could not find suitable data. It is not meant to be an authoritative model of data center energy consumption.

Hyper-scale cloud computing's 7 percent share of server stock is derived from an estimation that 5 percent of all global servers are owned by Google, Microsoft, Facebook, and eBay, and then extrapolated for other large hyper-scale cloud providers. There were 32 million total installed servers in 2010 (IDC installed base of servers from 2010) plus an estimated 1.3 million servers from self-assemblers like Google, Facebook, and others.[91,92]

NRDC's 2011 report *Small Server Rooms, Big Energy Savings* states that "large, mid-tier, and enterprise-class data centers comprise only half of all U.S. servers. The other half are housed in the small server rooms and closets typically found in small and medium businesses and organizations, as well as in departments and branch offices of larger organizations." Taking into account that some of the small- to medium-sized data centers have servers in the cloud slightly lowers the percentage estimate. Masanet et al., in the 2011 report *Estimating the Energy Use and Efficiency Potential of U.S. Data Centers*, used 2008 IDC data to estimate data center energy consumption by server type and data center space type. The SMO category comprises three data center types considered: "server closet," "server room," and "localized"; added together, their energy use equals 47 percent of the total.[93,94]

Total electricity consumption of U.S. data centers (76,400,000 MWh) is derived from the upper- and lower-bound average 2010 electricity consumption estimates for the United States.[95]

Using watts per server at an average utilization level from the SPECPower_ssj2008 benchmark provides for a normalization of energy consumption by data center type, while also including server age and PUE. Differences in efficiency between servers in hyper-scale cloud data centers versus small- to medium-sized data centers are addressed by using the 25th percentile of SPECPower data for hyper-scale and the 75th percentile for small to medium-size data centers.

The numbers in the table may not add up exactly due to rounding errors.

Endnotes

1 Using Jonathan Koomey, *Growth in Data Center Electricity Use 2005 to 2010*, analytics press, august 2011, http://www.analyticspress.com/datacenters.html for the 2010 baseline and emissions growth from Smarter 2020 (*GeSI SMARTer 2020: The Role of ICT in Driving a Sustainable Future*, Global e-Sustainability Initiative and the Boston Consulting Group, Inc., December 2012), as a proxy for electricity consumption growth. This is a low estimate because the carbon intensity of the U.S. grid is expected to be lower in 2020 than in 2010. These numbers are based on some of the most authoritative estimates available, however they have a high level of uncertainty.

2 DCD Intelligence, *Powering the Data Center*, Datacenter Dynamics, April 2013, www.dcd-intelligence.com/Products-Services/Powering-the-Datacenter.

3 Jonathan Koomey, *Growth in Data Center Electricity Use 2005 to 2010,* Analytics Press, August 2011, www.analyticspress.com/data centers.html.

4 R. Brown et al., *Report to Congress on Server and Data Center Energy Efficiency: Public Law 109@431*, Lawrence Berkeley National Laboratory, 2007.

5 Global e-Sustainability Initiative and the Boston Consulting Group, Inc., *GeSI SMARTer 2020: The Role of ICT in Driving a Sustainable Future*, December 2012.

6 Jonathan Koomey, "The NYT Article on Power, Pollution, and the Internet: My Initial Comments," September 25, 2012, www.koomey.com/post/32281701993.

7 Mirko Lorenz, *How Many Servers Worldwide?,* Vision Cloud, August 2011, www.visioncloud.eu/content.php?s=191,324.

8 David Freeland, *Colocation and Managed Hosting, Focus Investment Banking*, 2012, www.focusbankers.com/telecom/pdfs/FOCUS_Colocation_Winter12.pdf.

9 Josh Whitney, Phone Interview with 451 Research Director, November 2013.

10 Eric Masanet, Arman Shehabi, and Jonathan Koomey, "Characteristics of Low-Carbon Data Centers," *Nature Climate Change*, July 2013, vol. 3, no. 7, pp. 627-630, www.dx.doi.org/10.1038/nclimate1786 and www.nature.com/nclimate/journal/v3/n7/abs/nclimate1786.html#supplementary-information.

11 Tier 1 Research, "Multi-Tenant Data Center Category Definitions," presentation at Critical Facilities Roundtable meeting, November 6, 2009.

12 Uptime Institute, *2013 Uptime Institute Data Center Industry Survey*, http://uptimeinstitute.com/2013-survey-results (accessed November 2013).

13 Victor Avelar, Dan Azevedo, and Alan French, eds., *PUE: A Comprehensive Examination of the Metric*, The Green Grid, 2012.

14 Digital Realty Trust, *North American Campus Survey Results*, January 2013.

15 Uptime Institute, *2013 Uptime Institute Data Center Industry Survey*.

16 Kenneth Brill, *IT and Facilities Initiatives for Improved Data Center Energy Efficiency*, Uptime Institute, 2009.

17 Uptime Institute, *2013 Uptime Institute Data Center Industry Survey*.

18 451 Research, *DCIM Segment Focus: Market Overview Report*, April 2013.

19 Eric Masanet, Arman Shehabi, and Jonathan Koomey, "Characteristics of Low-Carbon Data Centers."

20 Bill Snyder, "Server Virtualization Has Stalled, Despite the Hype," *InfoWorld* online, December 31, 2010, www.infoworld.com/t/server-virtualization/what-you-missed-server-virtualization-has-stalled-despite-the-hype-901?source=footer.

21 Alex Benik, *The Sorry State of Server Utilization and the Impending Post-Hypervisor Era*, Gigaom, November 30, 2013, www.gigaom.com/2013/11/30/the-sorry-state-of-server-utilization-and-the-impending-post-hypervisor-era/.

22 Whitney et al., *The Carbon Emissions of Server Computing for Small- to Medium-Sized Organizations: A Performance Study of On-Premise vs. the Cloud*, NRDC and WSP Environment & Energy, October 2012.

23 Ibid.

24 L.A. Barroso and Urs Hölzle, *The Data Center as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, 2nd ed., Morgan & Claypool, 2013.

25 Paul Venezia, "Server Virtualization Deep Dive Report," *InfoWorld*, May 2011, www.infoworld.com/sites/infoworld.com/files/server virtualization deep dive.pdf.

26 Gartner Research, "Virtualization Penetration Rates in the Enterprise, 2012," *Cloud, IAAS & Virtualization Survey 2013*, Opsview Ltd., http://visual.ly/cloud-iaas-virtualization-adoption-2013.

27 Kim Mays, "SMB Adopting Virtualization with Success," *IT Business Edge*, August 30, 2013, www.itbusinessedge.com/blogs/smb-tech/smbs-adopting-virtualization-with-success.html.

28 Thomas Bittman, *Top 5 Virtualization Trends in 2012*, Gartner Research, March 21, 2012, www.blogs.gartner.com/thomas_bittman/2012/03/21/top-five-server-virtualization-trends-2012/.

29 VMware, *VMware Business and Financial Benefits of Virtualization*, 2011, www.vmware.com/files/pdf/cloud-journey/VMware-Business-Financial-Benefits-Virtualization-Whitepaper.pdf.

30 Sergey Blagodurov et al., *Maximizing Server Utilization While Meeting Critical SLAs via Weight-Based Collocation Management*, Integrated Network Management, 2013 IFIP/IEEE International Symposium, May 2013, Ghent, Belgium.

31 Scott Matteson, "10 Things You Shouldn't Virtualize," *TechRepublic*, August 8, 2013, www.techrepublic.com/blog/10-things/10-things-you-shouldnt-virtualize/.

32 Paul Rubens, "Top 10 Reasons Not to Virtualize," *IT Business Edge*, November 19, 2012, www.serverwatch.com/server-trends/top-10-reasons-not-to-virtualize.html.

33 Uptime Institute, *2013 Uptime Institute Data Center Industry Survey*.

34 Ibid.

35 Mark Blackburn, ed., *Unused Servers Survey Results Analysis*, The Green Grid, 2010.

36 Jonathan Koomey, "The NYT Article on Power, Pollution, and the Internet."

37 Mark Blackburn, ed., *Unused Servers Survey Results Analysis*.

38 Whitney et al., *The Carbon Emissions of Server Computing for Small- to Medium-Sized Organizations*.

39 Jonathan Koomey, Power-related Advantages of Cloud Computing, Uptime Institute, 2010.

40 *Salesforce.com* and the *Environment: Reducing Carbon Emissions in the Cloud*, Salesforce.com and WSP Environmental, 2010, www.salesforce.com/assets/pdf/misc/WP_WSP_Salesforce_Environment.pdf.

41 NRDC, *Is Cloud Computing Always Greener?* October 2012, www.nrdc.org/energy/files/cloud-computing-efficiency-IB.pdf.

42 Pierre Delforge, email exchange with John Clinger, ENERGY STAR Servers Program, July 2014.

43 Kenneth G. Brill and John Stanley, *IT and Facilities Initiatives for Improved Data Center Energy Efficiency*, Uptime Institute, September 2, 2009.

44 Standard Performance Evaluation Corp., SPECpower_ssj2008, www.spec.org/power_ssj2008/.

45 Kenneth Brill, "Shaving Millions Off Data Centers," *Forbes*, June 3, 2009, www.forbes.com/2009/06/02/energy-star-datacenter-technology-cio-network-energy-star.html.

46 Jonathan Koomey, *A Simple Model for Determining True Total Cost of Ownership for Data Centers*, Uptime Institute, 2007.

47 Eric Smalley, "IT Minds Quest for 'Holy Grail' of Data-Center Metrics," *Wired*, December 9, 2011, www.wired.com/wiredenterprise/2011/12/data-center-metric/.

48 The Green Grid, *The Green Grid Data Center Compute Efficiency Metric: DCcE*, 2010.

49 Jonathan Koomey, *The Economics of Green DRAM in Servers, Analytics Press*, November 2, 2012, www.mediafire.com/view/uj8j4ibos8cd9j3/Full_report_for_econ_of_green_RAM-v7.pdf.

50 Barbara Morris, Interview with the Expert, Nlyte Software Blog, 2011, www.blog.nlyte.com/interview-with-the-expert.

51 Uptime Institute, *2013 Uptime Institute Data Center Industry Survey*.

52 Jonathan Koomey, "The NYT Article on Power, Pollution, and the Internet."

53 Jonathan Koomey, *Four Reasons Why Cloud Computing Is More Efficient*, July 24, 2012, www.koomey.com/post/8014999803.

54 The Green Grid, *Breaking New Ground on Data Center Efficiency*, February 2012.

55 Nicole Schuetz, Anna Kovaleva, and Jonathan Koomey, eBay Inc.: *A Case Study of Organizational Change Underlying Technical Infrastructure Optimization*, Stanford University, September 26, 2013, www.law.stanford.edu/sites/default/files/publication/442226/doc/slspublic/Stanford%20eBay%20Case%20Study-%20FINAL-130926.pdf.

56 Gary Cullen et al., *Evaluation, Verification, and Measurement Study FY 2008/2009 Program for Silicon Valley Power*, Summit Blue Consulting, 2009, www.ncpa.com/images/stories/LegReg/SVP% percent20-% percent20EM&V% percent20Report% percent203-2-10% percent20final.pdf.

57 DCD Intelligence, 2013-2014 Census Report: North America Data Center Markets, Datacenters Dynamics, January 2014, www.dcd-intelligence.com.

58 David Freeland, *Colocation and Managed Hosting*.

59 451 Research, *Multi-Tenant Data Center Global Providers—2012*, November 2012.

60 Gary Cullen et al., *Evaluation, Verification, and Measurement Study*.

61 The Green Grid, *Breaking New Ground*.

62 Pierre Delforge, email exchange with John Clinger, ENERGY STAR Servers Program, July 2014.

63 Gary Cullen et al., *Evaluation, Verification, and Measurement Study*.

64 Ibid.

65 Rich Miller, *Data Center Leasing: It's All About the Megawatts*, Data Center Knowledge, November 11, 2009, www.datacenterknowledge.com/archives/2009/11/11/data-center-leasing-its-all-about-the-megawatts.

66 Katie Fehrenbacher, *10 Reasons Apple, Facebook & Google Chose North Carolina for Their Mega Data Centers*, Gigaom, July 10, 2012, www.gigaom.com/2012/07/10/10-reasons-apple-facebook-google-chose-north-carolina-for-their-mega-data-centers.

67 Patrick Thibodeau, "Apple, Google, Facebook Turn N.C. into Data Center Hub," *Computerworld* online, June 3, 2011, www.computerworld.com/s/article/9217259/Apple_Google_Facebook_turn_N.C._into_data_center_hub.

68 Gary Cullen et al., *Evaluation, Verification, and Measurement Study*.

69 New York State Energy Research and Development Authority, *Turnkey Internet & XAND—Data Center Case Study*, March 2013.

70 E.R. Masanet et al., "Estimating the Energy Use and Efficiency Potential of U.S. Data Centers," *Proceedings of the IEEE* 99, no. 8, pp 1440-1453, August 2011.

71 E.R. Masanet et al., "Estimating the Energy Use and Efficiency Potential of U.S. Data Centers."

72 The Green Grid, *Proxy Proposals for Measuring Data Center Efficiency*, January 5, 2009, http://www.thegreengrid.org/~/media/WhitePapers/White%20Paper%2017%20-%20Proxies%20Proposals%20for%20Measuring%20Data%20Center%20Efficiencyv2.pdf.

73 The Green Grid, *The Green Grid Data Center Compute Efficiency-Metric*, 2010.

74 Sustainability Accounting Standards Board, *Software & IT Services Sustainability Accounting Standard Exposure Draft 1*, 2013, www.sasb.org.

75 Eric Masanet, Arman Shehabi, and Jonathan Koomey, "Characteristics of Low-Carbon Data Centers."

76 John Bruschi et al., *FEMP Best Practices Guide for Energy-Efficient Data Center Design, U.S. Department of Energy*, Federal Energy Management Program, March 4, 2011, www1.eere.energy.gov/femp/pdfs/eedatacenterbestpractices.pdf.

77 Gary Cullen et al., *Evaluation, Verification, and Measurement Study*.

78 Joe Reele, "Data Centers & the Importance of Energy Efficiency in Co-location, Cloud and Hosted IT Environments," *Energy Manager Today*, December 11, 2013, www.energymanagertoday.com/data-centers-the-importance-of-energy-efficiency-in-co-location-cloud-and-hosted-it-environments-097531/?utm_source=el&utm_campaign=homefeed&utm_medium=link.

79 Tier 1 Research, "Multi-Tenant Data Center Category Definitions."

80 451 Research, *Multi-Tenant Data Center Global Providers*.

81 David Freeland, *Colocation and Managed Hosting*.

82 Ibid.

83 451 Research, *Multi-Tenant Data Center Global Providers*.

84 David Freeland, *Colocation and Managed Hosting*.

85 Gary Cullen et al., {*Evaluation, Verification, and Measurement Study FY 2008/2009 Program for Silicon Valley Power*, Summit Blue Consulting, 2009, www.ncpa.com/images/stories/LegReg/SVP%20-%20EM&V%.

86 David Freeland, *Colocation and Managed Hosting*.

87 451 Research, *Multi-Tenant Data Center Global Providers*.

88 David Freeland, *Colocation and Managed Hosting*.

89 Ibid.

90 451 Research, *Multi-Tenant Data Center Global Providers*.

91 Mirko Lorenz, *How Many Servers Worldwide?*

92 Jonathan Koomey, *Growth in Data Center Electricity Use*.

93 Drew Bennett and Pierre Delforge, *Small Server Rooms, Big Energy Savings*, NRDC, February 2012, www.nrdc.org/energy/files/Saving-Energy-Server-Rooms-IssuePaper.pdf.

94 E.R. Masanet et al., "Estimating the Energy Use and Efficiency Potential of U.S. Data Centers."

95 Jonathan Koomey, *Growth in Data Center Electricity Use.*

Natural Resources Defense Council
40 West 20th Street
New York, NY 10011
212 727-2700
Fax 212 727-1773

Beijing

Chicago

Los Angeles

Bozeman

San Francisco

Washington, D.C.

www.nrdc.org

www.nrdc.org/energy
www.twitter.com/nrdcEnergy
www.twitter.com/nrdc
www.facebook.com/nrdc.org

Printed on recycled paper