Title

Data center growth in the United States: decoupling the demand for services from electricity use

Permalink

https://escholarship.org/uc/item/5nh7f5vx

Journal

Environmental Research Letters, 13(12)

ISSN

1748-9318

Authors

Shehabi, Arman

Smith, Sarah J

Masanet, Eric

et al.

Publication Date

2018

DOI

10.1088/1748-9326/aaec9c

Peer reviewed

Data center growth in the United States: decoupling the demand for services from electricity use

Environmental Research Letters

OPEN ACCESS

RECEIVED
15 June 2018

REVISED
25 October 2018

ACCEPTED FOR PUBLICATION
30 October 2018

PUBLISHED
18 December 2018

LETTER

Data center growth in the United States: decoupling the demand for services from electricity use

Arman Shehabi[1], Sarah J Smith[1], Eric Masanet[2] and Jonathan Koomey[3]

[1] Energy Analysis & Environmental Impacts Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States of America
[2] McCormick School of Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, United States of America
[3] Koomey Analytics, PO Box 1545, Burlingame, CA 94011-1545, United States of America

E-mail: ashehabi@lbl.gov

Keywords: information and communication technology (ICT), energy efficiency, energy demand modeling, data centers

Supplementary material for this article is available online

Abstract

Data centers are energy intensive buildings that have grown in size and number to meet the increasing demands of a digital economy. This paper presents a bottom-up model to estimate data center electricity demand in the United States over a 20 year period and examines observed and projected electricity use trends in the context of changing data center operations. Results indicate a rapidly increasing electricity demand at the turn of the century that has significantly subsided to a nearly steady annual electricity use of about 70 billion kWh in recent years. While data center workloads continue to grow exponentially, comparable increases in electricity demand have been avoided through the adoption of key energy efficiency measures and a shift towards large cloud-based service providers. Alternative projections from the model illustrate the wide range in potential electricity that could be consumed to support data centers, with the US data center workload demand estimated for 2020 requiring a total electricity use that varies by about 135 billion kWh, depending on the adoption rate of efficiency measures during this decade. While recent improvements in data center energy efficiency have been a success, the growth of data center electricity use beyond 2020 is uncertain, as modeled trends indicate that the efficiency measures of the past may not be enough for the data center workloads of the future. The results show that successful stabilization of data center electricity will require new innovations in data center efficiency to further decouple electricity demand from the ever-growing demand for data center services.

Introduction

Data centers are the backbone of the information and communication technology that is becoming increasingly integral to our economy and society. Data center buildings house information technology (IT) equipment such a servers, storage and network equipment, as well as the infrastructure equipment needed to support IT electrical and thermal requirements. While an obscure building type 20 years ago, nearly all companies now employ some form of data center for their digital needs and these buildings are central to the services provided by companies in the growing and robust technology sector. As video streaming expands and the number of internet-connected devices continues to grow exponentially [1], data centers will be part of the supporting infrastructure needed to process, store, and transmit more and more zettabytes of data [2].

The high density of equipment in data centers makes them extremely energy intensive, often requiring 10–100 times more electricity per floor space area than other building types [3, 4]. Concern regarding the electricity demand from data centers, along with its impact on the electricity grid and broader energy sector, arose in the early 2000s as data centers rapidly proliferated to support the surge in digital services associated with widespread Internet access. Initial reports showed data center energy doubling from 2000 to 2005 both in the US and globally [5, 6]. Facing such rapid growth and the potential for overwhelming electricity demand from data centers, the US Congress requested a report that ultimately estimated that US data centers had consumed about 61 billion kilowatt-hours (kWh) in 2006 (1.6% of total US electricity sales) for a total electricity cost of about $4.5 billion (2006 dollars) [7].

The Report to Congress (Public Law 109-431 [8]), led to a bottom-up modeling framework, outlined in Masanet *et al* (2011) [9], that drew from earlier studies to create a reproducible model and allowed users to compare projected impacts of US electricity demand under different scenarios for data center design and operation. An additional study using a similar methodology estimated US data center electricity use had grown to about 2% of total US electricity sales in 2010, but noted a decrease in the rise of electricity demand in 2008 and 2009, which was primarily attributed to the economic recession [10].

The growth in data center energy demand observed in these studies led to speculation that US data center energy use would pass 100 billion kWh before 2020 [11], but in 2016 the US Department of Energy (DOE) issued a report that showed a surprising reduction in US data center energy growth since 2010 [12], though still representing approximately 36% of global data center energy use in 2014 [13]. The DOE report was developed in anticipation of additional congressional requests [14] and provides estimates of US data center energy use through the year 2020 using an expanded modeling framework that accounts for changes that have occurred in the data center sector since the previous studies, most notably the prevalence of cloud computing and the rise of large 'hyperscale' data centers.

This paper provides further insight into to the unexpected trends generated by the model and discusses how US data center electricity use may continue to change beyond 2020. Historical and projected trends are examined in the context of the changing data center workload demand and energy efficiency implementation. Two alternative scenarios for the 2010–2020 decade are presented to illustrate the wide range in potential electricity use needed to support data centers and the role of energy efficiency in decoupling electricity demand from data center growth. Additionally, a new metric is proposed—the full processor equivalent (FPE)—to quantify the energy intensity of per-processor trends in computing and data center efficiency, as well as highlight the relationship between the demand for services and the corresponding electricity requirements in future growth projections of the data center industry. Finally, this paper also documents the mathematical framework of the model used by DOE's 2016 report, providing a reproducible and expandable version of the model that can be refined when new data become available and altered to account for any future changes in the data center sector. A detailed description of the mathematical framework of the model, including indexed calculations for each equipment and space type, is presented in the supplemental online material (SOM), available online at stacks.iop.org/ERL/13/124030/mmedia.

Modeling methodology and assumptions

Data center space types

The data center energy model utilizes a bottom-up approach with equipment-level estimates in order to estimate electricity use. Electricity use (E) is modeled as the sum of electricity use of four equipment categories (servers (E^S), storage (E^{ST}), network (E^P), and infrastructure (E^I)) (equation (1)) across eleven data center space types based on widely-used taxonomy from the International Data Corporation (IDC) [15]. These space types span six sizes: room, closet, localized, mid-tier, enterprise, and hyperscale, as well as two usage types: internal and service provider. Internal data centers represent traditional facilities that support businesses and institutions, while service provider data centers account for specialized facilities that represent the core services of businesses such as communication and social media companies. Under this taxonomy, service provider data centers also include colocation facilities, where space within a data center is leased to businesses that procure and manage their own IT equipment [16]. The six size categories have distinctive infrastructure and operational characteristics as described in Shehabi *et al* [11]. The largest size, hyperscale, represents a relatively new segment of warehouse-size facilities that have emerged with the growth in cloud platforms, mobile devices, social media, and big data. Hyperscale data centers tend to operate more efficiently in terms of IT equipment use (e.g. higher server utilizations) [6, 17, 18], as well as their infrastructure systems (e.g. more efficient building cooling designs) [19–21]. Additionally, this is a rapidly growing data center category, with some firms estimating that 53% of all servers will be in hyperscale datacenters by 2021 [22]

$$E = E^S + E^{ST} + E^P + E^I. \tag{1}$$

Scenario overview

The model is used to estimate data center energy use across the entire United States in three scenarios. The 'Current Trends' scenario couples historical and projected equipment shipments with expected baseline improvements in equipment efficiency and operational practices from 2000 to 2020. This estimate of data center energy use is contrasted against two alternative scenarios for the years 2010–2020 to illustrate the range in possible data center energy demand over that decade that would be attributable to the implementation of energy efficiency practices. The 'Frozen Efficiency' alternative holds energy efficiency practices at 2010 levels while the increases in demand for data center services and server computational improvements continue to match current trends through 2020. Lastly, beginning in 2010, the 'Best Practices' alternative assumes widespread adoption by 2020 of the most efficient technologies and best management practices applicable to each data center type while, again, the data

Table 1. Modeled 2010 historic US data center characteristics and projected 2020 characteristics under three different efficiency scenarios.

Variable	Units	2010	2020 Current trends	2020 Frozen efficiency	2020 Best practices
Baseline server installed base	million	14.3	18.3	35.4	10.2
Server wattage	W				
1-socket volume servers					
Maximum		118	118	118	118
Minimum (idle)		70	48	70	33
2-socket + volume servers					
Maximum		365	365	365	365
Minimum (idle)		216	149	216	103
Volume server utilization	%				
Internal datacenter		10	15	10	32
Service provider datacenter		20	25	20	45
Hyperscale datacenter		40	50	45	70
Volume server utilization	%				
Weighted average		14	28	14	52
Server average wattage	W				
1-socket volume servers		75	58	75	62
2-socket + volume servers		238	213	238	246
Mid-range servers		1200	1880	1880	1880
High-end servers		13700	20200	20200	20200
Storage capacity installed base	million TB				
HDD		32.2	665	665	665
SSD		0.9	292	292	292
Storage drive capacity	TB/drive				
HDD		0.9	10.0	10.0	10.0
SSD		0.2	5.0	5.0	5.0
Storage average wattage	W/drive				
HDD		11.3	6.5	11.3	6.5
SSD		6.0	6.0	6.0	6.0
Network port installed base	million	40.2	87.4	87.4	57.2
Network port wattage	W/port				
100 Mb		1.6	0.6	1.6	0.5
1000 Mb		2.6	1.0	2.6	0.8
10 Gb		4.1	1.6	4.1	1.2
40 Gb		7.0	2.7	7.0	2.0
PUE					
Closets		2.00	2.00	2.00	2.0
Rooms		2.60	2.35	2.60	1.5
Localized		2.08	1.88	2.08	1.5
Mid-tier		1.98	1.79	1.98	1.4
High-end		1.77	1.60	1.77	1.3
Hyperscale		1.25	1.13	1.25	1.1
PUE					
Weighted average		1.9	1.51	1.9	1.25

center service demand and server computations continue to match current trends through 2020. Key assumptions for both alternative scenarios, as well as the current trend inputs are summarized in table 1. Data and assumptions for each scenario are derived from literature, industry data, and expert solicitation compiled in Shehabi *et al* [11]. All applied data and assumptions, as well model equations and intermediate calculated values are presented in the SOM.

Servers

Categorization of servers

Servers represent the most significant use of energy in data centers. As in Masanet *et al* [8], the model adopts the IDC categorization of volume, midrange, and high-end servers. In this model, volume servers are further disaggregated into four categories based on the number of processor sockets they contain (1-socket or 2 or more sockets) and the type of vendor from which they were purchased (branded or unbranded). Grouping servers by socket count improves accuracy in estimating the wattage of servers, as 1-socket (1S) servers use considerably less energy than the more prevalent 2-socket (2S+) type [23]. For vendor type, 'branded' represents traditional supply chains where servers are designed and sold through large original equipment manufacturers (e.g. Hewlett–Packard, Dell), while 'unbranded' refers to a newer business model where servers are made to customer specifications and sold directly from the original design manufacturer (ODM). Though the model assumes branded and unbranded servers have identical energy use characteristics, maintaining the separation in vendor type provides a proxy for the server count in hyperscale service provider data centers through the use of industry data that tracks the ODM server market, as unbranded servers are almost exclusively sold to this data center space type [24–26].

Server installed base

The total installed base of each type of server, as well as the total server count in each of the eleven space types, are inputs to the model and based on data from IDC's Worldwide Quarterly Server tracker [27]. Mid-range and high-end servers are distributed across the space types based on an assumed distribution (see SOM), while volume servers are distributed by assuming that all unbranded servers are located in hyperscale datacenters and that the ratio of 1S to 2S+ servers is constant across all space types. This server distribution creates a baseline server count for every server type and space type combination, which is then modified to become the actual estimated server count (N^S) based on the implementation of efficiency measures, namely the removal of servers that are no longer being used ('inactive' servers) and consolidation of less-utilized servers onto fewer, higher-utilized machines.

Volume servers are by far the most common server type, representing more than 95% of the US server installed base. Volume servers fall into three operational categories: inactive ($N^{S,I}$), active consolidated ($N^{S,C}$), and active non-consolidated ($N^{S,A}$). Inactive servers (also referred to as comatose or 'zombie' servers), represent obsolete or unused servers that consume electricity but provide no useful information services. Previous studies have estimated that inactive servers represent 10%–30% of servers in US data centers [28–31]. Removal of these servers is an opportunity to save energy, and highlights the impact of raised awareness on the part of data center operators as to what equipment is being used in the data center. In this analysis, inactive servers are conservatively assumed to make up 10% and 5% of baseline volume servers in internal and service provider data centers, respectively, so as not to overestimate the potential savings from their removal. The Current Trends and Frozen Efficiency scenarios assume inactive servers stay constant at these percentages over time. The Best Practices scenario assumes the fraction of inactive servers removed through efficiency efforts grows linearly from zero–one (total removal) from 2010–2020.

For active servers, a key efficiency opportunity is consolidation, which entails replacing multiple servers running at low processor utilization (non-consolidated) with a single server running at a higher processor utilization (consolidated), using methods such as virtualization and containerization [16]. The Current Trends scenario inherently includes some consolidation, as represented in IDC forecasts and increasing utilization assumptions. No additional consolidation occurs in this scenario. The Frozen Efficiency scenario removes this inherent consolidation by assuming utilization stays frozen at 2010 levels. However, workload demand for data center services still increases identically to the Current Trends scenario, therefore requiring additional servers in the installed base to provide the same amount of overall computation at a lower per-server utilization level. In the Best Practices scenario, 80% of baseline active volume servers are consolidated by 2020 onto servers that run at high utilization levels of 45% for internal data centers, 55% for non-hyperscale service provider data centers, and 75% for hyperscale datacenters.

When consolidating servers, 'overhead' utilization occurs due to applications that must be run on the server to balance multiple workloads. This analysis assumes 'overhead' utilization increases the utilization of virtualized servers by 5% [11]. For example, if two servers previously running at 10% utilization were consolidated to one server, and the utilization overhead was 5%, the resulting server would need to run at 25% utilization. The specific assumptions and equations involved in estimating the count of consolidated and non-consolidated servers are detailed in the SOM.

Once the number of inactive, active consolidated, and active non-consolidated volume servers are estimated, they are aggregated to the total server count of each volume server type (i) in each space type (j), as shown in equation (2)

$$N_{ij}^S = N_{ij}^{S,A} + N_{ij}^{S,C} + N_{ij}^{S,I}. \qquad (2)$$

Electricity use

The number of servers estimated in the installed base, as described above, is multiplied by the average per-sever electricity use (e^S) to calculate total server energy use (E^S) in each year (equation (3)). Power draw for mid-range and high-end servers is estimated at an average level across the installed base. Midrange servers are estimated to consume approximately 890 W in 2014 and 1880 W in 2020, while high-end servers are estimated to consume 10 600 and 20 200 W in those years, based on the assumptions outlined in Shehabi *et al* [12], with roughly linear growth between values

$$E_{ij}^S = N_{ij}^S \, e_{ij}^S. \qquad (3)$$

Volume server electricity use is modeled using a baseline maximum (e_{max}^S) and idle (e_{idle}^S) energy use, a graphics processing unit (GPU) scaling factor (g), and utilization (u) (equation (4)). Maximum wattage for 1S and 2S+ volume servers was estimated from the Server Efficiency Rating Tool (SERT)[4] database as 118 W and 365 W respectively [21]. These power

[4] SERT was created by SPEC for the US Environmental Protection Agency's (EPA) ENERGY STAR program. This tool uses a set of synthetic worklets to test discrete system components, providing detailed power use data at different load levels. Data from this tool is submitted to the EPA by manufacturers, and is collected and maintained by the Information Technology Industry Council (ITI). Data collected by ITI through March 2016 was used in this report.

estimates correspond to an overall weighted volume server average maximum wattage of ~330 W, which is consistent with previous work [6]. Temporally constant maximum power is also observed in the Standard Performance Evaluation Corporation's (SPEC) SPEC Power database[5,6], which shows approximately constant maximum power in servers from 2007 to 2015 [32], as well as other previous studies [33]. Therefore, this analysis assumes these wattages are constant from 2010 to 2020. Additionally, branded and unbranded servers with the same socket count are assumed to have the same maximum power

$$e^S_{ij} = (e^S_{max\,,\,j} - e^S_{idle,j})g_{ij}u_i + e^S_{idle,j}g_i. \tag{4}$$

Idle power use is estimated based on an assumed ratio of idle power to maximum power, referred to as the dynamic range. Reducing this ratio is a key efficiency opportunity for servers, which generally operate at low utilization levels [34]. The dynamic range is assumed to be the same across volume server types and decreases over time as servers become more efficient. Idle power is assumed, according to Shehabi *et al* [11], to be approximately 60% of maximum power in 2010 and to be about 40% and 30% of maximum power in 2020 for the Current Trends and Best Practices scenarios, respectively. Idle power remains 60% of maximum power in 2020 (the same as 2010) for the Frozen Efficiency scenario.

While the potential growth of GPUs in servers has received increased attention with the emergence computational methods such as machine learning, GPU-powered servers still constitute just a fraction of the server stock, with only about 5% of global server shipments including any GPUs in 2016 [35]. In this analysis, no change in server energy use is assumed from GPUs through 2020 due to their currently low representation in servers and the lack of data regarding future adoption and energy impacts. However, the potential growth in GPU use for a wide array of emerging applications [36] contributes to the uncertainty in long-range projections of annual global data center traffic that vary by nearly 80 zettabytes by 2030 and drive global data center use estimates as high as 8 PWh per year [37]. Consequently, the GPU scaling factor remains in the model to emphasize that estimates of GPU penetration in the server stock should continue to be monitored and revisited in future analyses of server power use.

Lastly, the average utilization level for servers is calculated as the weighted average of the utilization of active non-consolidated servers, active consolidated servers, and inactive servers. Inactive servers have utilization of 0, while consolidated servers operate at the utilizations discussed in the previous section. Utilization for active non-consolidated volume servers varies by space type, and linearly increases from 2010 –2020 to account for the growing level of virtualization in data centers. Service provider data centers are assumed to run at higher utilizations than internal data centers, as the servers in service provider data centers are often configured for more specialized and predictable operations. Hyperscale data centers are assumed to run at higher utilizations than other service providers and internal data centers based on estimates in cloud and non-cloud data centers [6, 15, 16].

Storage

Data center storage is disaggregated between hard disk drive (HDD) and solid state drive (SSD) technologies, due to differences in energy usage between the two types. The storage installed base, in terms of terabyte (TB) capacity, is based on data from IDC's Worldwide Quarterly Storage Tracker [38] and represents storage drives in external devices separate from servers, as well as any drives internal to servers with three or more drives installed. The first two storage drives within a server are not considered in the storage installed base, as the energy use of those drives are assumed on average to be captured in the server energy use metrics. The capacity of the installed base is distributed across space types assuming (1) no storage (beyond the first two internal server drives) is present in server rooms and closets and (2) storage is present in the remaining space types in proportion to the number of servers present. Storage capacity is then converted to number of drives (N^{ST}) for each drive type (k) using per-drive capacity assumptions from Shehabi *et al* [12]: 0.9 TB/drive in 2010 and 10 TB/drive in 2020 for HDD, and 200 GB/drive and 5 TB/drive for SSD in 2010 and 2020, respectively. Conversion to per-drive values is due to the availability of per-drive wattage estimates in the literature.

Storage electricity use (E^{ST}, equation (5)) is calculated using assumed baseline wattages (e^{ST}_{base}) for each storage type (k): 11.3 and 6.5 W/disk in 2010 and 2020 for HDD, and constant 6 W/drive for SSD, as the improvements in drive efficiency have typically been coupled with large increases in capacity [39, 40]. An additional operational energy factor (O) is assumed for drives in external devices; equal to 25% of the energy required for the storage drive itself [11]. Drives in external devices are estimated to account for 73% and 76% (F) of the storage installed base in 2014 and 2020, based on IDC shipment data [35].

The best practices scenario assumes the efficiency (n^{ST}) of both HDD and SSD storage systems improve linearly, beyond the 2010 baseline wattage, by 25% in 2020. Storage efficiency can be achieved by employing measures such as more efficient disk drive

[5] The SPEC Power benchmark suite measures power and performance of servers. SPECpower_ssj2008 is an industry-standard benchmark application that has been used since 2007, with users self-submitting results to a database that is reviewed and released to the public quarterly. Data through 2015 Q4 was used in this study.

[6] While the wattages reported in the SPEC database were not used directly due to the assumed self-selection bias towards high efficiency servers in the database, the general temporal trends are assumed to be representative of all servers.

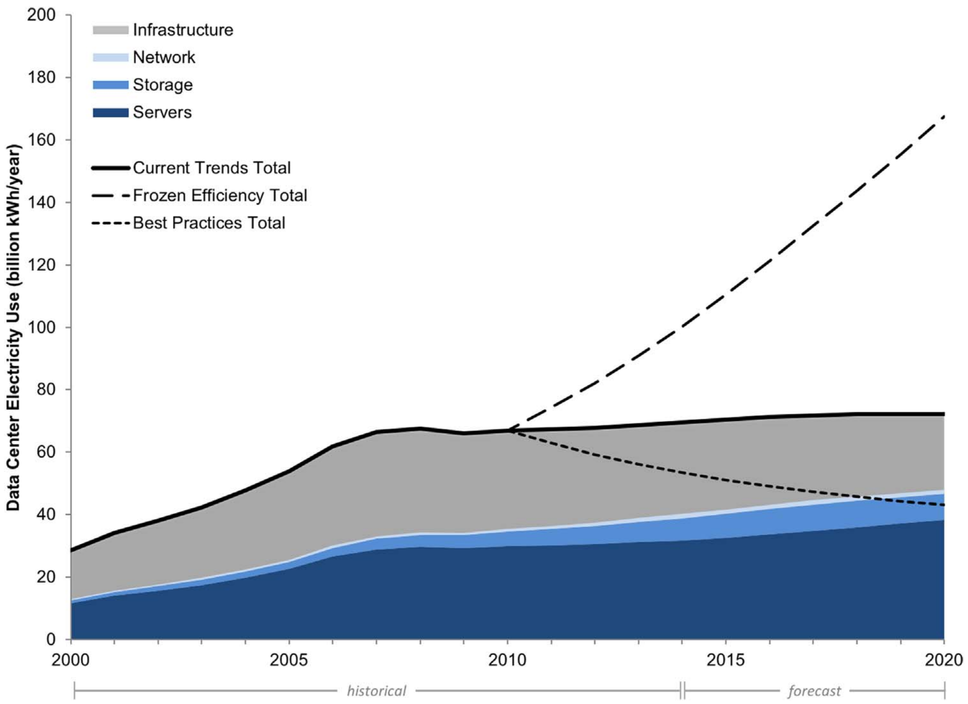


Figure 1. Current trends of US data center equipment electricity use from 2000–2020 (historical and projected), with two alternative scenarios beginning in 2010.

components, lower power use in idle states, and use of capacity optimization methods [37]

$$E_{ik}^{ST} = N_{ik}^{ST} e_{base,k}^{ST}(1 - n^{ST})(1 + O * F_k). \qquad (5)$$

Networking equipment

Energy use required for the transmission of data across the internal data center network (N^P) is estimated by modeling the electricity use of Level 2/3 networking ports inside data centers, as shown in equation (6). The model estimates network energy for four different port speeds (l): 100 MB, 1000 MB, 10 GB, and 40 GB, based on equipment shipment data from IDC's Worldwide Quarterly Network tracker [41]. Total network port estimates are distributed among space types in direct proportion to the number of servers in the given space type (i). While total number of ports per server is constant across the space categories, faster speed ports are weighted towards larger space categories, using the distribution methods described in the SOM. In the Best Practices scenario, baseline values of port counts are adjusted to account for network port consolidation measures, an efficiency opportunity similar to server consolidation, where 80% of 10 GB network ports are consolidated 4-to-1 into 40 GB ports by 2020. The final port count estimate (N^P) is then used in electricity calculations.

Baseline port wattage (e_{base}^P) is assumed to decrease linearly over time, based on previously published port wattages [7, 42], as well as a survey of 51 technical specification sheets followed by industry review [12]. 2010 values of 1.6, 2.6, 4.1, and 7.0 W are assumed for the four speeds, respectively, and decreasing to 0.6, 1.0, 1.6, and 2.7 W by 2020. The Best Practices scenario assumes all port speeds improve in efficiency (n^P) from 0%–25% 2010–2020. Average network port efficiency can be achieved using measures such as improvements in network topology, dynamic link rate adaptation, and link and switch sleep modes [43]

$$E_{i,l}^P = N_{i,l}^P e_{base,l}^P(1 - n^P). \qquad (6)$$

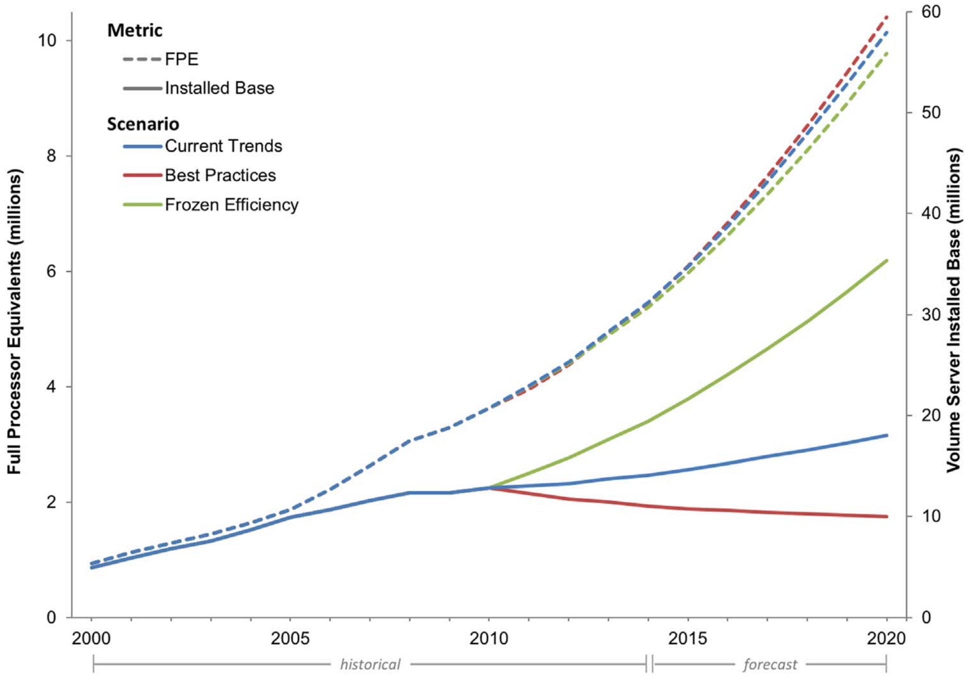
Infrastructure

Infrastructure energy use is calculated using the power usage effectiveness (PUE) metric [44]. In the context of this study, 'infrastructure' consists of the data center equipment that is not used solely for the purpose of performing computations or for the storage or transmission of data, such as cooling systems, lighting, and power supplies. The PUE metric represents total data center energy use relative to IT equipment energy use; e.g. for a PUE of 2, every watt of power used to power IT equipment results in an additional watt of infrastructure energy use. Therefore, infrastructure electricity use is calculated according to equation (7). Space type-specific PUE values for 2010–2020 are assumed for each scenario according to Shehabi *et al* [11] and presented in table 1

$$E_i^I = (E_i^S + E_i^{ST} + E_i^P)(\text{PUE}_i - 1). \qquad (7)$$

Results and discussion

Figure 1 presents modeled estimates of total US data center electricity use over a two-decade period, with



Note: overhead from virtualization use creates slight variations (<5%) among each scenario's FPE.

Figure 2. US volume server installed base and corresponding full-processor equivalent, 2000–2020.

estimates prior to 2010 using historical data and inputs from previous studies [6, 8, 11] and the 2010–2020 estimates based on the equipment tracking data and industry-validated efficiency trends described in the previous section. Electricity demand increases from about 29 billion kWh in 2000 to nearly 73 billion kWh by 2020, with most of the increase occurring during the first decade. From 2000–2005 electricity use nearly doubled to 56 billion kWh; a rapid increase that has been cited in previous studies [4, 5]. Electricity demand from 2005–2010 grew less dramatically, with an overall increase of 24%, which is clearly influenced by the conspicuous 2009 drop in electricity demand in accordance with the 2008 economic recession. Only a slight growth in data center electricity returns after the recession and this modest growth rate is expected to continue through 2020, resulting in just over 5% of an increase total electricity demand over the entire decade.

The fairly stable electricity demand estimated post-recession from 2010 through 2020 belies the influence of efficiency measures implemented over that same period. Figure 1 highlights the wide range in total data center energy use that results depending on the level of implementation of those efficiency measures through the two alternative scenarios. The Frozen Efficiency and Best Practices scenarios show total data center electricity use reaching drastically different levels over time, varying by nearly a factor of four by 2020, while maintaining the same workload demand for data center services and the same server computational improvements as in the Current Trends scenario. Electricity use in the Current Trends scenario is markedly lower than in the Frozen Efficiency scenario—suggesting great gains in data center energy efficiency since 2010—though major energy savings still remain untapped, as evidenced by the Best Practices scenario.

The demand for data center services in a specific year and the corresponding computational performance of server stock is represented by the FPE metric, which accounts for both the number of processors operating in volume servers and the average utilization of those processors. For example, 150 1-socket servers all running at 10% utilization would be represented by an FPE of 15 (i.e. equivalent to 15 processors running at 100% utilization). The FPE values in figure 2 represent the total number of processors in the US volume server stock, as well as the utilization of those processors which depends on data center type and the operational practices for the given year. Note that the FPE is only a metric of physical processor use and does not represent the quality or quantity of the computations that occur within that processor. Rather, the FPE estimated for a given year is simply a rough proxy of the computational demand relative to the installed processor stock for that specific year. Figure 2 shows that FPE nearly doubles from 2010–2017, but given that computational power of computer chips has historically increased exponentially [45], the 2017 stock of processors represented by an FPE of 7.9 million would have an order of magnitude more computational demand than the 2010 stock of processors represented by an FPE of 4.1 million.

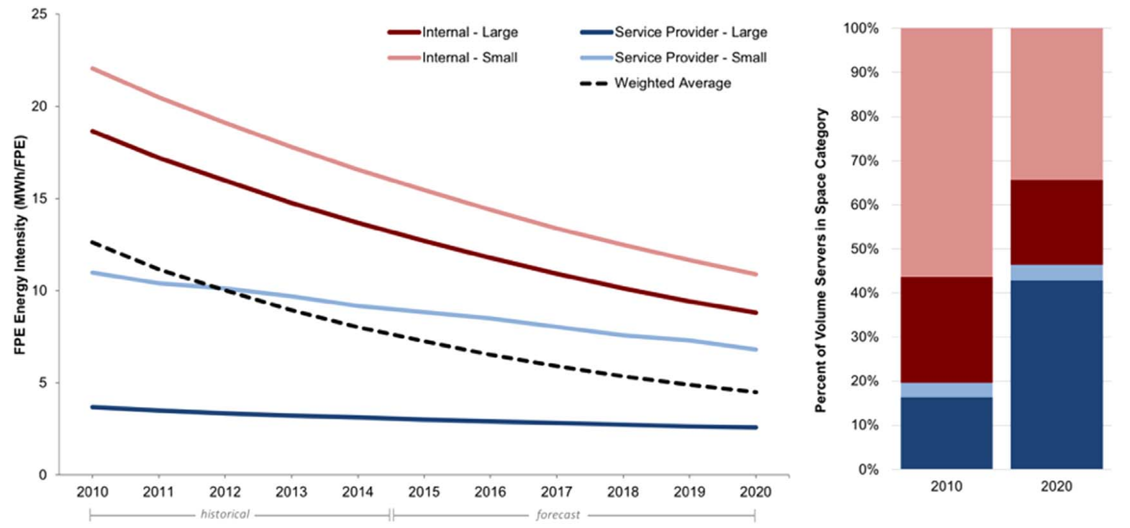


Figure 3. (a) Ratio of electricity use to total full processor equivalents for volume servers in four data center space categories, 2010–2020 (left), and (b) distribution of server installed base across four space categories, 2010 and 2020 (right). Data centers with floor space greater than 1850 m^2 (enterprise and hyperscale size categories) are represented as large.

In figure 2, note that the server installed based growth for the Current Trends, Frozen Efficiency, and Best Practices scenarios somewhat match the growth in data center electricity use for the corresponding scenarios in figure 1. The exponential growth of FPE in figure 2 for all three scenarios, however, shows that neither the installed base or electricity use are necessarily indicative of the workload demand for data center services. Rather, the periods of steady electricity demand in the Current Trends scenarios shown in figure 1 occur in the face of a corresponding FPE growth that more closely resembles the electricity use in the Frozen Efficiency scenario. This apparent decoupling of data center service output and electricity use is influenced by the market shift towards larger, more efficient, data centers.

Figure 3(a) presents a ratio of electricity use and FPE, defined here as the 'FPE energy intensity,' which represents the total electricity required to fully utilize the equivalent of one single volume server processor, both in terms of server operation and the associated infrastructure electricity. The FPE energy intensity improves over time for all data center types as efficiency measures are increasingly implemented, but significant variation in efficiency exists among the different space types. Larger data centers operated by service providers are generally more efficient, owing to economies-of-scale design advantages over smaller data centers, such as implementing cooling system economizers, and optimization strategies often unavailable to internal data centers, such as consolidating specialized and predictable operations. Figure 3(a) shows large service provider data centers have an FPE energy intensity nearly seven times lower than small internal data centers in 2010. The rapid emergence of hyperscale data centers, caused by demand for cloud computing, large-scale colocation, and the growth of service provider companies, has increased the portion of the installed processor stock operating in these large buildings, as shown in figure 3(b). This shift toward hyperscale has accelerated the improvement in the average FPE energy intensity of volume server processors in US data centers during this decade.

The Frozen Efficiency scenario in figure 1 shows that the energy impact of an improved average FPE energy intensity across the US data center stock has been significant. With FPE energy intensity remaining at 2010 levels in the alternative scenario, while FPE demand continues to grow at the exponential rate shown in figure 2, total data center electricity use increases to nearly 170 billion kWh annually by 2020, more than double the amount estimated in the Current Trends scenario. The Current Trends' improvement in FPE energy intensity relative to the Frozen Efficiency amounts to an accumulative savings across the decade (2010–2020) of more than 475 billion kWh; equivalent to the annual electricity use of 50 million households [46].

The Best Practices scenario in figure 1 highlights that additional savings are still available, with data center electricity use at only 45 billion kWh 2020; nearly half of the 72 billion kWh projected with Current Trends. The efficiency measures to achieve these Best Practices savings only include strategies that are already employed on a large scale, such as consolidation efforts to increase server utilization and cooling designs that reduce facility PUE. As with the Frozen Efficiency scenario, the Best Practices scenario does not consider computational improvements in CPUs, such as processing speed, which are still assumed to advance at the same rate as in Current Trends. The overall FPE demand remains essentially the same in all three scenarios, by design, with only slight (<5%) variations due to differences in server virtualization adoption and the corresponding utilization overhead.

While the Current Trends and Best Practices scenario estimates in this analysis show the significant electricity savings available from the adoption of known efficiency measures, the contradicting trends in figures 2 and 3 indicate that the recent stability in electricity demand may be a limited phenomenon. As more and more of the data center stock is represented by the most efficient data center types, the potential for known improvements diminish, thus slowing the rate of improving the FPE energy intensity. A slowing rate of energy efficiency improvement in the face of exponential FPE demand portends the potential return to growing electricity needs in the data center sector.

Ultimately, the future growth in this sector's electricity use is dependent on the balance of data center demand (represented as FPE) and energy efficiency (represented as FPE energy intensity), where forecasts of either variables contain high levels of uncertainty in a rapidly evolving sector that is known for disruption. In terms of data center demand, historical exponential FPE growth may underestimate the future data center needs from an emerging internet-of-things economy [47] or from the potential increase in GPU use to support autonomous vehicles and other services associated with artificial intelligence [48]. FPE growth may also accelerate from a slowing of Moore's Law [49], a previously highlighted concern [50, 51] where additional processors beyond historical observation might be needed to meet the continued growth of computational demand. The slowing of Moore's Law could have a significant impact on data center energy use and has already been estimated to begin slowing the rate of efficiency improvements in processors by 2022, causing the projected increases in global energy use to roughly double by 2030 [36]. Alternatively, the FPE growth rate could slow if significant breakthroughs in computing cause future utilized processors to do much more computational work than what is expected from Moore's Law (e.g. quantum computing), requiring fewer processors to provide the same services.

In terms of data center energy efficiency, future improvements in FPE energy intensity are dependent on the adoption rate of known efficiency measures as well as the development of new efficiency opportunities. The rate of improvement in efficiency can be expected to slow as the implementation of known efficiency measures continue to shift the average FPE energy intensity of the data center stock closer to that of the best hyperscale data centers, which operate at maximum utilizations and PUEs nearing unity. Major innovations in data center design, however, could potentially drop the power required to operate data centers below current conceptual limits.

Conclusion

Data center energy use modeling is a challenging endeavor given the rapid evolution of digital services, the quick turnover of IT equipment stock, and the proprietary nature of this economic sector. This paper provides updates and presents insight into to the unexpected trends generated by the 2016 DOE data center model. The FPE metric is introduced to capture the relationship between data center demand and energy efficiency implementation over time and across different data center types. Two alternative scenarios are also presented to highlight how energy efficiency can help decouple electricity demand from the demand for data center services and how further improvements are available with known efficiency measures. Finally, this paper also documents the DOE data center model structure, allowing for future energy impact comparisons between different technologies and practices to help identify pathways toward lower energy demand.

Model results of three scenarios presented highlight the significant impact of efficiency measures, with nearly the same estimated data center demand (expressed in FPE) for 2020 requiring a national electricity use that varies by about 135 billion kWh. This wide range in electricity use required to support a given demand of processor utilization shows the impact of certain energy efficiency opportunities that improve power scaling, increase processor utilization, and reduce PUE, all of which have significantly improved across the US data center stock since 2010. These improvements have also been accelerated by the market growth of large service provider data centers (i.e. hyperscale) that are often attentively operated at high utilizations in buildings with efficiently designed cooling systems. Additionally, cloud computing and colocation have provided an alternative to the small inefficient data centers that typically contain underutilized servers and inefficient cooling.

The trend in data center electricity use since 2000 is a success story of energy efficiency. Rapidly increasing electricity demand at the turn of the century led to the development and implementation of innovative energy efficiency strategies that curbed electricity growth while data center demand continued to grow exponentially. The growth of data center electricity use beyond 2020, however, is uncertain as the modeled trends indicate efficiency measures of the past my not be enough for the data center demand of the future, further highlighting the need for new innovations in data center efficiency to be developed and implemented at a rate consummate with the ever-growing demand for digital services from these buildings.

Acknowledgments

This material is based upon work conducted by Lawrence Berkeley National Laboratory with support from the Department of Energy Advanced Manufacturing Office. Lawrence Berkeley National Laboratory is supported by the Office of Science of the United States Department of Energy and operated under Contract Grant No. DE-AC02-05CH11231.

ORCID iDs

Arman Shehabi https://orcid.org/0000-0002-1735-6973
Sarah J Smith https://orcid.org/0000-0003-0179-4546
Jonathan Koomey https://orcid.org/0000-0002-2983-344X

References

[1] CVN Index 2017 The Zettabyte Era: Trends and Analysis (Cisco white paper)

[2] Jones N 2018 How to stop data centres from gobbling up the world's electricity *Nature* **561** 163

[3] Greenberg S, Mills E, Tschudi W, Rumsey P and Myatt B 2006 Best practices for data centers: results from benchmarking 22 data centers *Proc. 2006 ACEEE Summer Study on Energy Efficiency in Buildings (Asilomar, CA)*

[4] Capuccio D and Craver L 2007 The data center power and cooling challenge (The Gartner Group)

[5] Koomey J G 2007 Estimating Total Power Consumption by Servers in the US and the World (15 February)

[6] Koomey J 2008 Worldwide electricity used in data centers *Environ. Res. Lett.* **3** 034008

[7] Brown R 2007 *Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431: Appendices LBNL-363E* Lawrence Berkeley National Laboratory (Berkeley, CA) (https://doi.org/10.2172/929724)

[8] Code of Federal Regulations, an Act to Study and Promote the Use of Energy Efficient Computer Servers in the United States (Public Law 109–431, 120 STAT. 2920, 2006)

[9] Masanet E, Brown R E, Shehabi A, Koomey J G and Nordman B 2011 Estimating the energy use and efficiency potential of US data centers *Proc. IEEE* **99** 1440–53

[10] Koomey J G 2011 *Growth in Data Center Electricity Use 2005 to 2010* (Oakland, CA: Analytics Press)

[11] Delforge P and Whitney J 2014 Issue paper: data center efficiency assessment scaling up energy efficiency across the data center industry: evaluating key drivers and barriers Natural Resources Defense Council (NRDC)

[12] Shehabi A, Smith S J, Horner N, Azevedo I, Brown R, Koomey J, Masanet E, Sartor D, Herrlin M and Lintner W 2016 *United States data center energy usage report LBNL-1005775* Lawrence Berkeley National Laboratory (Berkeley, CA) (https://doi.org/10.2172/1372902)

[13] International Energy Agency 2017 Digitalization & Energy. Organization for Economic Co-operation and Development (Paris, France)

[14] 115th Congress 2017 S. 1460: Energy and Natural Resources Act of 2017 (GovTrack) (31 July 2017) (www.govtrack.us/congress/bills/115/s1460)

[15] Villars R L 2014 'US Datacenter Census and Construction 2014–2018 Forecast: Realigning Workloads, Managing Obsolescence, and Leveraging Hyperscale' (IDC #252712)

[16] Chen N, Ren X, Ren S and Wierman A 2015 Greening multi-tenant data center demand response *Perform. Eval.* **91** 229–54

[17] NRDC and WSP 2012 The Carbon Emissions of Server Computing for Small- to Medium-Sized Organizations: A Performance Study of On-Premise versus The Cloud. WSP Environment & Energy, LLC and Natural Resources Defense Council (October 2012)

[18] Barroso L A, Clidaras J and Hölzle U 2013 *The Datacenter as a Computer: An Introduction to the Design of Warehouse-scale Machines* (*Synthesis Lectures on Computer Architecture*) (San Rafael, CA: Morgan Claypool Publishers) (https://doi.org/10.2200/S00516ED2V01Y201306CAC024)

[19] Google 2015 *Our energy-saving data centers* (Accessed: 2 December 2015) (http://google.com/about/datacenters/efficiency/internal/index.html#measuring-efficiency)

[20] Gelber R 2012 *Facebook showcases green datacenter. HPCwire* (http://hpcwire.com/hpcwire/2012-04-26/facebook_showcases_green_datacenter.html) (Accessed: 2 December 2015)

[21] Masanet E, Shehabi A, Ramakrishnan L, Liang J, Ma X, Walker B and Mantha P 2013 *The energy efficiency potential of cloud-based software: a US case study LBNL-6298E* Lawrence Berkeley National Laboratory (Berkeley, CA) (https://doi.org/10.2172/1171159)

[22] Cisco 2018 Cisco Global Cloud Index: Forecast and Methodology (2016–2021)

[23] Dietrich J 2014 ITIC Analysis of SERT Worklet Results (Information Technology Industry Council)

[24] TCP 2014 ODM servers to see explosive growth in coming years: IDC (The China Post) (14 March 2014)

[25] Gartner 2014 Gartner Says Data Center Infrastructure ODMs Are a Key Threat to Data Center OEMs' Direct Business (Gartner Newsroom Press Release) (3 September 2014)

[26] Pietroforte M 2013 ODM Direct Servers (cloud) market boosts while overall server sales decline (4Sysops) (11 December 2013)

[27] 2015 International Data Corporation (IDC) IDC's Worldwide Quarterly Server Shipment Tracker (2010–2018, Framingham, MA, March)

[28] Kaplan J M, Forrest W and Kindler N 2008 Revolutionizing Data Center Efficiency McKinsey and Company

[29] The Uptime Institute estimate (https://uptimeinstitute.com/research-publications/asset/comatose-server-savings-calculator) (Accessed: 9 November 2018)

[30] Koomey J and Taylor J 2015 New data supports finding that 30 percent of servers are 'Comatose', indicating that nearly a third of capital in enterprise data centers is wasted (TSO logic)

[31] McMillian R 2015 Zombie servers: they're here and doing nothing but burning energy (The Wall Street Journal) (13 September 2015)

[32] SPEC 2015 SPECpower_ssj2008 Results (10 September 2015) (https://spec.org/power_ssj2008/results/)

[33] Van Heddeghem W, Lambert S, Lannoo B, Colle D, Pickavet M and Demeester P 2014 Trends in worldwide ICT electricity consumption from 2007 to 2012 *Comput. Commun.* **50** 64–76

[34] Fuchs H, Shehabi A, Ganeshalingam M, Desroches L B, Lim B, Roth K and Tsao A 2017 *Characteristics and energy use of volume servers in the United States LBNL-2001074* Lawrence Berkeley National Laboratory (Berkeley, CA) (https://doi.org/10.2172/1350977)

[35] International Data Corporation (IDC) 2017 Personal communication with Peter Rutten and Lidice Fernandez (3 May 2017)

[36] Dean J, Patterson D and Young C 2018 A new golden age in computer architecture: empowering the machine learning revolution *IEEE Micro* **38** 21–9

[37] Andrae A S and Edler T 2015 On global electricity usage of communication technology: trends to 2030 *Challenges* **6** 117–57

[38] International Data Corporation (IDC) 2015 IDC's Worldwide Quarterly Disk Storage Systems Tracker 2010–2019 (Framingham, MA, March)

[39] ASHRAE 2015 Data Center Storage Equipment—Thermal Guidelines, Issues, and Best Practices Technical Committee 9.9

[40] Reinsel D 2010 A Plateau in Sight for the Rising Costs to Power and Cool the World's External Storage? IDC Opinion, IDC#225016. September 2010

[41] International Data Corporation (IDC) 2015 IDC's Worldwide Quarterly Data Center Networks 2008–2019, Framingham, MA, March

[42] Lanzisera S, Nordman B and Brown R E 2012 Data network equipment energy use and savings potential in buildings *Energy Efficiency* **5** 149–62

[43] Dudkowski D and Hasselmeyer P 2015 Energy-efficient networking in modern data centers *Green Communications: Principles, Concepts and Practice* ed K Samdanis *et al* (New York: Wiley)

[44] Belady C, Rawson A, Pfleuger J and Cader T The Green Grid datacenter power efficiency metrics: PUE and DCiE (Technical Report)

[45] Koomey J G, Berard S, Sanchez M and Wong H 2011 Implications of historical trends in the electrical efficiency of computing *IEEE Ann. Hist. Comput.* **33** 46–54

[46] EPA 2017 Greenhouse Gas Equivalencies Calculator. United States Environmental Protection Agency (https://epa.gov/energy/greenhouse-gas-equivalencies-calculator) (Accessed: 9 November 2018)

[47] Scarbrough D 2017 Data centers are gearing up to harness IoT tech (Data Center Dynamics)

[48] MSV J 2017 In The Era Of Artificial Intelligence, GPUs Are The New CPUs (Forbes Media, LLC)

[49] Koomey J and Naffziger S 2016 Energy efficiency of computing: what's next? In Electronic Design (28 November)

[50] Bashroush R 2018 A comprehensive reasoning framework for hardware refresh in data centres *IEEE Trans. Sustain. Comput.* (accepted) (https://doi.org/10.1109/TSUSC.2018.2795465)

[51] Malmodin J and Lundén D 2018 The energy and carbon footprint of the global ICT and E&M sectors 2010–2015 *Sustainability* **10** 3027