POLICY FORUM

As demand for data centers rises, energy efficiency improvements to the IT devices and cooling systems they house can keep energy use in check.

PHOTO: GOOGLE

ENERGY

Recalibrating global data center energy-use estimates

Growth in energy use has slowed owing to efficiency gains that smart policies can help maintain in the near term

By **Eric Masanet[1,2], Arman Shehabi[3], Nuoa Lei[1], Sarah Smith[3], Jonathan Koomey[4]**

Data centers represent the information backbone of an increasingly digitalized world. Demand for their services has been rising rapidly (*1*), and data-intensive technologies such as artificial intelligence, smart and connected energy systems, distributed manufacturing systems, and autonomous vehicles promise to increase demand further (*2*). Given that data centers are energy-intensive enterprises, estimated to account for around 1% of worldwide electricity use, these trends have clear implications for global energy demand and must be analyzed rigorously. Several oft-cited yet simplistic analyses claim that the energy used by the world's data centers has doubled over the past decade and that their energy use will triple or even quadruple within the next decade (*3*–*5*). Such estimates contribute to a conventional wisdom (*5*, *6*) that as demand for data center services rises rapidly, so too must their global energy use. But such extrapolations based on recent service demand growth indicators overlook strong countervailing energy efficiency trends that have occurred in parallel (see the first figure). Here, we integrate new data from different sources that have emerged recently and suggest more modest growth in global data center energy use (see the second figure). This provides policy-makers and energy analysts a recalibrated understanding of global data center energy use, its drivers, and near-term efficiency potential.

Assessing implications of growing demand for data centers requires robust understanding of the scale and drivers of global data center energy use that has eluded many policy-makers and energy analysts. The reason for this blind spot is a historical lack of "bottom-up" information on data center types and locations, their information technology (IT) equipment, and their energy efficiency trends. This has led to a sporadic and often contradictory literature on global data center energy use.

Understanding where data center energy use is heading requires considering service demand growth factors alongside myriad equipment, energy efficiency, and market structure factors (see the first figure). Bottom-up analyses tend to best reflect this broad range of factors, generating the most credible historical and near-term energy-use estimates (*7*). Despite several recent national studies (*8*), the latest fully replicable bottom-up estimates of global data center energy use appeared nearly a decade ago. These estimates suggested that the worldwide energy use of data centers had grown from 153 terawatt-hours (TWh) in 2005 to between 203 and 273 TWh by 2010, totaling 1.1 to 1.5% of global electricity use (*9*).

Since 2010, however, the data center landscape has changed dramatically (see the first figure). By 2018, global data center workloads and compute instances had increased more than sixfold, whereas data center internet protocol (IP) traffic had increased by more than 10-fold (*1*). Data center storage capacity has also grown rapidly, increasing by an estimated factor of 25 over the same time period (*1*, *8*). There has been a tendency among analysts to use such service demand trends to simply extrapolate earlier bottom-up energy values, leading to unreliable predictions of current and future global data center energy use (*3*–*5*). They might, for example, scale up previous bottom-up values (e.g., total data center energy use in 2010) on the basis of the growth rate of a service demand indicator (e.g., growth in global IP traffic from 2010 to 2020) to arrive at an estimate of future energy use (e.g., total data center energy use in 2020).

But since 2010, electricity use per computation of a typical volume server—the workhorse of the data center—has dropped by a factor of four, largely owing to processor-efficiency improvements and reductions in idle power (*10*). At the same time, the watts per terabyte of installed storage has dropped by an estimated factor of nine owing to storage-drive density and efficiency gains (*8*). Furthermore, growth in the number of servers has slowed considerably owing to a fivefold increase in the average number of compute instances hosted per server (owing to virtualization), alongside steady reductions in data center power usage effectiveness (PUE, the total amount of energy used by a data center divided by the energy used by its IT equipment). Both of these trends have been largely driven by shifts in compute instances to energy-efficient cloud and "hyperscale" data centers, the largest data center type (*1*, *2*). In the United States—the world's largest data center market—industry-vetted bottom-up analyses of these efficiency trends identified a plateau in national data center en-

[1]McCormick School of Engineering and Applied Science, Northwestern University, Evanston, IL, USA. [2]Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, USA. [3]Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [4]Koomey Analytics, Burlingame, CA, USA.
Email: eric.masanet@northwestern.edu

ergy use since 2010, despite rapid increases in demand for U.S. data center services (*11*). We now expand that analysis to the global level and show that strong continued efficiency progress can maintain an energy use plateau for the next few years through proactive policy initiatives and data center energy-management practices. These new bottom-up estimates form the basis of recent global data center energy values utilized by the International Energy Agency (*12*).

The data leveraged here facilitate a more technology-rich and temporally consistent approach than was available previously. Since 2011, analysts at Cisco have published data and outlooks for worldwide server stocks, data center workloads, server virtualization levels, and storage estimates for traditional, cloud, and, most recently, hyperscale data centers (*1*). In a series of reports starting in 2016, Lawrence Berkeley National Laboratory has published energy trend analyses of servers, storage devices, and network devices commonly used within data centers (*8*, *11*, *13*). Analysts have documented the numbers and locations of hyperscale data centers that represent a substantial fraction of global data center compute instances, and major data center operators are increasingly reporting their PUE (*14*).

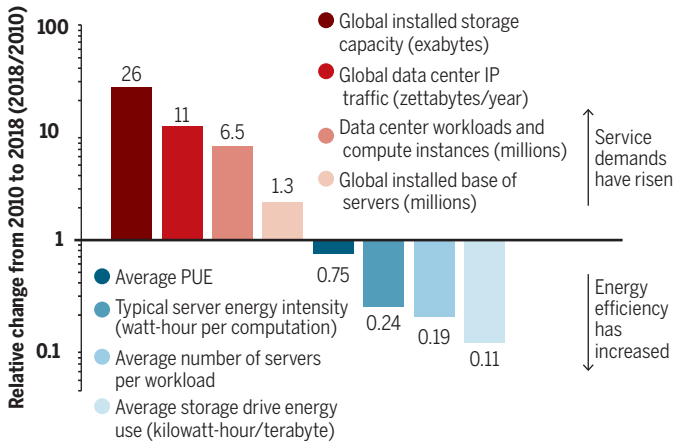
When integrated into a bottom-up modeling framework, these data suggest that, although global data center energy has increased slightly since 2010, growth in energy use has been substantially decoupled from growth in data center compute instances over the same time period (see the second figure, second graph). Moreover, the refined view provided by these new data suggests that global data center energy use in 2010 was around 194 TWh, slightly less than the lower-bound estimate in the 2010 bottom-up study (203 TWh) when fewer data were available (*9*).

In 2018, we estimated that global data center energy use rose to 205 TWh, or around 1% of global electricity consumption. This represents a 6% increase compared with 2010, whereas global data center compute instances increased by 550% over the same time period. Expressed as energy use per compute instance, the energy intensity of global data centers has decreased by 20% annually since 2010, a notable improvement compared with recent annual efficiency gains in other major demand sectors (e.g., aviation and industry), which are an order of magnitude lower (*12*).

The new integrated data illuminate some key technological and structural trends that help explain these large energy intensity improvements (see the first figure and the second figure, second graph). The combination of increased server efficiencies and greater server virtualization (which reduces the amount of server power required for each compute instance) has enabled a sixfold increase in compute instances with only a 25% increase in global server energy use, whereas the combination of increased storage-drive efficiencies and densities has enabled a 25-fold increase in storage capacity with only a threefold increase in global storage energy use. Shifts to faster and more energy-efficient port technologies have enabled a 10-fold increase in data center IP traffic with only modest increases in network device energy use. In sum, although overall energy use of IT devices (servers, storage, and network) has increased from around 92 TWh in 2010 to around 130 TWh in 2018, technological and operational efficiency gains have enabled substantial growth in services with comparatively much smaller growth in energy use.

Notably, the new data also suggest a large decrease in the energy use of data center infrastructure systems (i.e., cooling and power provisioning), enough to mostly offset the growth in total IT device energy use. This decrease is explainable by ongoing shifts in servers away from smaller traditional data centers (79% of compute instances in 2010) and toward larger and more energy-efficient cloud (including hyperscale) data centers (89% of compute instances in 2018) (see the second figure, third graph), which have much lower reported PUE values owing to cutting-edge cooling-system and power-supply efficiencies (*1*, *11*).

Trends in global data center energy-use drivers



PUE, power usage effectiveness; IP, internet protocol.

Yet given ever-growing demand for data center services, how much longer can these current efficiency trends last? Predicting the long-term efficiency limits of IT devices is notoriously difficult, especially in light of potential game-changing technologies such as quantum computing, for which energy use is unclear (*2*). Yet over the near term, market analysts predict that even greater levels of server virtualization are feasible (*1*), and technology studies indicate remaining potential for IT device efficiency gains, including more shifts to low-power storage devices (*8*). On the infrastructure side, world-class hyperscale data centers are already operating with PUEs of 1.1 or lower, which is close to the practical minimum value. Additional structural shifts from smaller traditional data centers to hyperscale data centers are predicted in the near term (*1*), indicating that infrastructure energy use may be dampened even further. Should these trends play out over the next few years, our approach indicates that there is a sufficient energy efficiency resource to absorb the next doubling of data center compute instances that would occur in parallel with a negligible increase in global data center energy use (see the second figure, second graph).

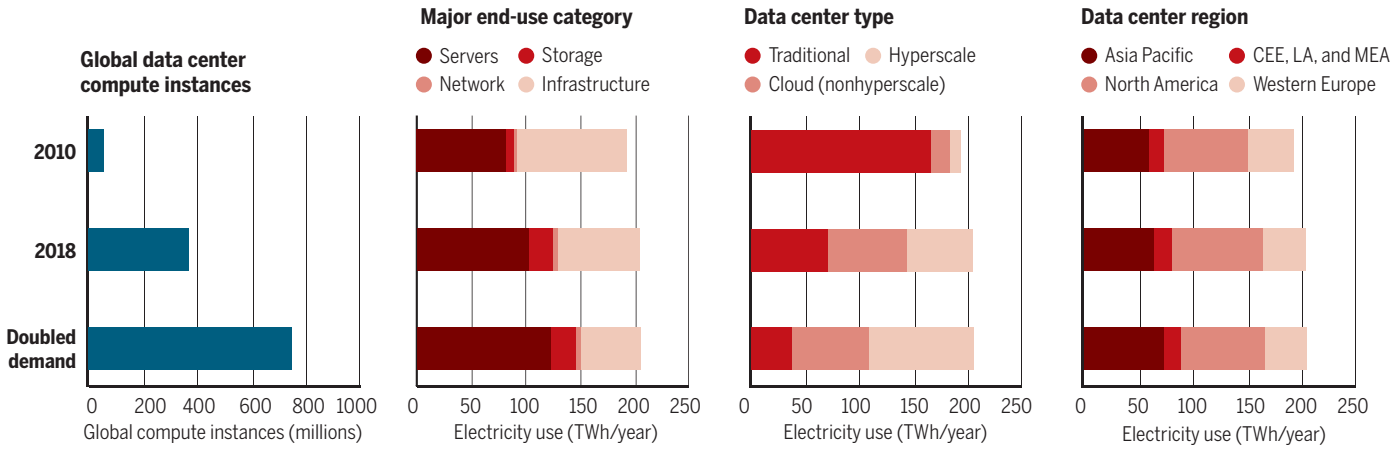
These findings lie in contrast to recent predictions of rapid and unavoidable near-term energy demand growth. Yet the IT industry, data center operators, and policy-makers can't rest on their laurels; diligent efforts will be required to manage possibly sharp energy demand growth once the existing efficiency resource is fully tapped. The next doubling of global data center compute instances may occur within the next 3 to 4 years (*1*).

For policy-makers, there are three main areas of action. First, policy support can help data centers seize the remaining efficiency potential of current technology and structural trends. One key strategy includes further strengthening and promotion of efficiency standards such as Energy Star for servers, storage, and network devices while requiring such certifications in public IT procurement programs. Efficiency standards give data center operators access to more efficient IT devices while creating strong market incentives to manufacturers to continue innovating energy-efficient products. To support such standards, greater investments are needed to develop energy efficiency benchmarks for storage and network devices—similar to the Standard Performance Evaluation Corporation's (SPEC's) SPEC Power bench-

GRAPHIC: X. LIU/*SCIENCE*

Historical energy usage and projected energy usage under doubled computing demand

Doubled demand (relative to 2018) reflects current efficiency trends continuing alongside predicted growth in compute instances.



CEE, LA, and MEA, Central and Eastern Europe, Latin America, and Middle East and Africa; TWh, terrawatt-hour.

mark for servers—while policy should require that measured performance of all certified IT devices be made public to spur ongoing competition. Another strategy is to incentivize shifts to cloud services when economically and institutionally feasible—for example, through procurement standards and utility rebates—ensuring that future compute instances are delivered by data centers at the cutting edge of energy efficiency. Yet another is to encourage and incentivize continuous reductions in PUE, some of which are attainable through low-cost measures such as improved airflow management and temperature set-point optimization and through vehicles such as subsidized energy efficiency audits and tax credits. These and other proven data center efficiency strategies (*2*, *7*, *8*) can bring about a near-term plateau in energy use, which provides critical time to prepare for the possibility of future energy demand growth. But this time must be used wisely.

Second, investment in new technologies is needed to manage future energy demand growth in the cleanest manner possible once current efficiency trends reach their feasible limits. Strong deployment incentives should be provided to accelerate the pace of renewable energy adoption by data centers, including low-carbon procurement standards and corporate tax credits, so that the carbon intensity of current and future energy demand is reduced substantially (*15*). And greater public funding should be allocated to advancements in computing, data storage, communications, and heat removal technologies that may extend the IT industry's historical efficiency gains well into the future. Key examples include quantum computing, materials for ultrahigh density storage, increased chip specialization, artificial intelligence for computing resource and infrastructure management, and liquid and immersion cooling technologies. However, it is crucial to increase investments immediately to ensure such technologies are economical and scalable in time to prevent a demand surge later this decade, which would also make required renewable capacity additions more challenging.

Third, much greater public data and modeling capacities are required for understanding and monitoring data center energy use and its drivers and for designing and evaluating effective policies. National policy-makers should enact robust data collection and open data repository systems for data center energy use, in much the same way as has been done historically for other demand sectors. Proprietary data concerns can be addressed through data reporting and aggregation protocols, similar to energy data for the industrial sector, which shares many of the same confidentiality concerns (see, for example, the U.S. Manufacturing Energy Consumption Survey). Such efforts are important in all world regions and particularly in Asia, where data center energy use is poised to grow (see the second figure, fourth graph), but reliable data are scarce, especially for China, where data centers are multiplying quickly. In parallel, more public reporting by large data center operators should be encouraged and incentivized (e.g., through efficiency rating systems) for greater energy-use transparency and accountability.

To make full use of these important data, more research funding is needed for developing policy-relevant data center energy models and for model sharing and research community building that can disseminate and ensure best analytical practices. With better data, analysts should also quantify uncertainties in future modeling results, leading to more robust policy decisions. Given the important role data centers will play in future energy systems, the historical dearth of knowledge on their energy use and the mixed signals given to policy-makers by contradictory findings are unacceptable. Global data center energy use is entering a critical transition phase; to ensure a low-carbon and energy-efficient future, we cannot wait another decade for the next reliable bottom-up estimates. ■

REFERENCES AND NOTES

1. Cisco, "Cisco Global Cloud Index: Forecast and methodology, 2016–2021 white paper" (Cisco, document 1513879861264127, 2018).
2. International Energy Agency (IEA), *Digitalization & Energy* (IEA, 2017).
3. L. Belkhir, A. Elmeligi, *J. Clean. Prod.* **177**, 448 (2018).
4. A.S.G. Andrae, T. Edler, *Challenges* **6**, 117 (2015).
5. T. Bawdy, "Global warming: Data centres to consume three times as much energy in next decade, experts warn," *The Independent*, 23 January 2016.
6. N. Jones, *Nature* **561**, 163 (2018).
7. E. Masanet, R. E. Brown, A. Shehabi, J. G. Koomey, B. Nordman, *Proc. IEEE* **99**, 1440 (2011).
8. A. Shehabi *et al.*, "United States data center energy usage report" (Lawrence Berkeley National Laboratory, LBNL-1005775, 2016).
9. J. G. Koomey, "Growth in data center electricity use 2005 to 2010" (Analytics Press for the *New York Times*, 2011).
10. B. Wagner, "Intergenerational energy efficiency of Dell EMC PowerEdge servers" (Dell, DellEMC white paper, 2018).
11. A. Shehabi, S. J. Smith, E. Masanet, J. Koomey, *Environ. Res. Lett.* **13**, 124030 (2018).
12. IEA, "Tracking clean energy progress" (IEA, 2019); www.iea.org/tcep/.
13. H. Fuchs *et al.*, *Energy Effic.* 10.1007/s12053-019-09809-8 (2019).
14. M. Avgerinou, P. Bertoldi, L. Castellazzi, *Energies* **10**, 1470 (2017).
15. E. Masanet, A. Shehabi, J. G. Koomey, *Nat. Clim. Chang.* **3**, 627 (2013).

ACKNOWLEDGMENTS

This material includes work conducted by Lawrence Berkeley National Laboratory (LBNL) with support from the U.S. Department of Energy (DOE) Advanced Manufacturing Office. LBNL is supported by the Office of Science of the DOE and operated under contract grant No. DE-AC02-05CH11231. E.M. and N.L. are grateful for financial support provided by Leslie and Mac McQuown. The global data center analysis modeling file with all data inputs, results, methodological notes, figures, discussion of uncertainties, and sources is available on GitHub (doi: 10.5281/zenodo.3668743).

10.1126/science.aba3758

GRAPHIC: X. LIU/*SCIENCE*

Recalibrating global data center energy-use estimates

Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith and Jonathan Koomey

Science **367** (6481), 984-986.
DOI: 10.1126/science.aba3758