Title
Expert and operator perspectives on barriers to energy efficiency in data centers

Journal
Energy Efficiency, 17(6)

ISSN
1570-646X

Authors
Newkirk, Alex C
Hanus, Nichole
Payne, Christopher T

Publication Date
2024-08-01

DOI
10.1007/s12053-024-10244-7

Energy Efficiency (2024) 17:63
https://doi.org/10.1007/s12053-024-10244-7

ORIGINAL ARTICLE

Expert and operator perspectives on barriers to energy efficiency in data centers

Alex C. Newkirk · Nichole Hanus · Christopher T. Payne

Received: 17 April 2024 / Accepted: 3 July 2024 / Published online: 17 July 2024

Abstract It was last estimated in 2016 that data centers (DCs) comprise approximately 2% of total US electricity consumption. However, this estimate is currently being updated to account for the massive increase in computing needs due to streaming, cryptocurrency, and artificial intelligence (AI). To prevent energy consumption that tracks with increasing computing needs, it is imperative we identify energy efficiency strategies and investments beyond the low-hanging fruit solutions. In a two-phased research approach, we ask: What non-technical barriers still impede energy efficiency (EE) practices and investments in the data center sector, and what can be done to overcome these barriers? In particular, we are focused on social and organizational barriers to EE. In Phase I, we performed a literature review and found that technical solutions are abundant in the literature, but fail to address the top-down cultural shifts that need to take place in order to adapt new energy efficiency strategies. In Phase II, reported here, we interviewed 16 data center operators/experts to ground-truth our literature findings. Our interview protocols focus on three aspects of DC decision-making: procurement practices, metrics and monitoring, and perceived barriers to energy efficiency. We find that vendors are the key drivers of procurement decisions, advanced efficiency metrics are facility-specific, and there is convergence in the design of advanced facilities due to the heat density of parallelized infrastructure. Our ultimate goals for our research are to design DC decarbonization policies that target organizational structure, empower individual staff, and foster a supportive external market.

Keywords Data centers · Energy efficiency · Organizational behavior · Non-technical barriers

A. C. Newkirk (✉) · N. Hanus · C. T. Payne
Building Technology and Urban Systems division, Lawrence Berkeley National Laboratory (LBNL), 1 Cyclotron Road, Berkeley, CA 94720, USA
e-mail: ACNewkirk@lbl.gov

A. C. Newkirk
Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA

Introduction

The best measure of data center (DC) energy consumption in the U.S. comes from a 2016 report from Lawrence Berkeley National Laboratory (LBNL), which estimates that DCs consume 2–5% of our electricity (Shehabi et al., 2016). LBNL is currently updating this model with new data surrounding workload and equipment types that represent more accurate load profiles for a proliferation of new technologies: cryptocurrency, ubiquitous streaming, machine learning (ML), and artificial intelligence (AI). Indeed, the U.S. dominates data center hardware shipments and needs and some estimates demonstrate that the U.S. DC industry is poised to grow at a combined annual growth rate (CAGR) of 14.2% between

2022 and 2032 (Niall McCarthy, 2021; Precedence Research, 2023). Without a strong commitment to advancing energy efficient practices and investments, DC energy consumption will likely grow with increased computing needs (Shehabi et al., 2018).

In a two-phased research approach, we ask: What non-technical barriers still impede energy efficiency (EE) practices and investments in the data center sector, and what can be done to overcome these barriers? In particular, we are focused on social and organizational barriers to EE. In Phase I (Hanus et al., 2023), we performed a literature review of 86 resources, ranging from peer-reviewed journal publications to handbooks. We also considered related fields such as organizational behavioral management and energy-intensive buildings. We found that technological solutions, although abundant in the literature, fall short of providing practical guidance on overcoming social and organizational barriers to EE. In fact, little is understood about how individual psychologies of information technology (IT) and facilities staff affect EE implementation in DCs. However, setting goals and providing incentives from the top down (flowing from the executive level of organizations to IT and facilities) can effect cultural change.

In Phase II, the phase we report on in this paper, we interviewed 16 data center operators/experts to ground-truth our literature findings. The remainder of this paper is organized as such: Methodology, Results, Discussion, and Conclusion. First, we describe how we developed our interview protocol, designed and recruited our sample, and performed quantitative and qualitative analysis on the resulting interview data. Next, we detail our interview results by organizing our findings into the three main categories contained in our interview protocol: procurement processes, metrics and monitoring, and barriers to energy efficiency perceived by our participants. We discuss the limitations of our work, a few policy implications, and next steps. Our ultimate goals for our research are to design DC decarbonization policies that target organizational structure, empower individual staff, and foster a supportive external market.

Methodology

Overview

This analysis is the second step in a two-phased research program. Phase II, presented here, constitutes an interview study to examine three main topics and explore how they differ from the literature: data center technology procurement and operations, performance tracking and metrics, and barriers to energy efficiency. It is difficult to reach data center owners/operators due to their limited resources (e.g., time and building staff) as well as their proclivity to protect proprietary information. Therefore, we decided to first employ an interview study with a convenience sample before employing a larger survey study or beginning a series of focus groups and workshops. Since we obtained a convenience sample, we do not make statistical claims of these findings. However, our interviews did allow us to explore the various factors that are important to energy efficiency investment and operations. Furthermore, our interviews allowed us to explore how participant perspectives differed from those already identified in our Phase I literature review (Hanus et al., 2023).

Interview protocol

The Human Subjects Committee at Lawrence Berkeley National Laboratory approved an exempt protocol for these interviews (Protocol ID Pro00023259 and approval number 407NR001-30MR23).

Our interview protocol was informed by the literature (Phase I) and the Mental Models approach (Morgan et al., 2001). We used a semi-structured format to allow interviewees to expound on topics to cover areas not yet identified in the literature or to underscore their importance. Based on our previous literature review, we organized our interview protocol into three sections: (1) data center investment and operation processes (e.g., "How is equipment specified?"), (2) data center performance measures and metering (e.g., "Please select all performance characteristics your organization uses to measure your data center performance."), and (3) barriers to energy efficiency (e.g., "From this list of common barriers, please select all that apply to your organization."). These three areas were of particular interest to the authors, as they speak to the differences in personal and organizational psychology of owning and operating a data center. After understanding these areas better, we can begin to provide practical guidance on avoiding the common pitfalls of implementation that the literature currently does not address.

We developed two versions of the interview protocol: one for data center experts/researchers (E/R and one for data center owners/operators (O/O).

Following the interview approach in Hanus et al. (2018), the content and structure of the protocols were similar to identify potentially important factors influencing data center EE decision-making between experts who are studying and consulting on data center operations, and individuals or organizations who actually own or operate them. Comparing these two groups is particularly useful for determining existing knowledge and context differences. The protocols were pilot tested in May 2022 for comprehensiveness and understandability with one operator at Lawrence Berkeley National Laboratory and one operator at University of California, Davis. Ultimately, our interview protocols aimed to answer the following three research questions:

i. What new barriers were identified in the interviews that were not found in the literature?
ii. What new metrics were identified in the interviews that were not found in the literature?
iii. How do perceptions of barriers to energy efficiency differ among experts vs. owners/operators and across data center types?

Recruitment and participants

We interviewed a total of 16 data center experts and owners/operators, and we leaned heavily on owners/operators (nine operator participants compared to seven experts). We began recruitment in earnest in June 2022 and sent 18 emails to prospective participants within our networks: the Center of Expertise for Energy Efficiency in Data Centers, the Data Center Energy Practitioner Training program, and Open Compute Project. Follow-up emails were sent in the event of a non-response and 16 participants ultimately agreed to an interview. During our first set of interviews, we employed snowball sampling – a convenience sampling method whereby participants suggest additional contacts who may be interested in participating in the study (Berg & Lune, 2011, p. 20). Of the 16 participants, three were female and 13 were male. Eight of the participants were from the federal/public sector (e.g., hospitals, national labs), and the remaining half were from the commercial sector. Four participants represented high-performance computing (HPC) data centers and one participant represented a hyperscale cloud provider. See Appendix 2 for more detailed participant characteristics. The interviews lasted approximately one-hour and interviewees were given the option for compensation via a $50 Visa gift card.

We categorized our owner/operator participants according to the type of facility with which they were most familiar. Operators responsible for multiple data centers we instructed to consider the one they had greatest involvement in. Facilities which provide IT services for the internal operations of their organization without an independent stream of revenue we consider "enterprise data centers". Facilities whose primary revenue stream is leasing access to their power and cooling infrastructure to external clients we consider "colocation (colo) data centers". Scientific and high-performance computing centers we initially categorized separate from enterprise cloud, but found over the course of our analysis that these overlapped significantly in both IT hardware, cooling infrastructure, and energy performance. To reflect this overlap, we created an aggregate category of "hyperscale" to describe any facility with 20 megawatts of installed IT load or greater. These categories possess some overlaps and edge cases, enterprise data centers can often act as "internal colos" for example. Our facility categorizations then reflect the efficiency challenges unique to different types of DC facilities identified in our previous research: the split incentive in the colocation business model and the operational competency of the largest facilities.

Analyses

All interviews were conducted in Zoom, audio-recorded, and transcribed. The interview protocols were coded and implemented in Qualtrics, with each interview involving one administrator and one note-taker. Therefore, the interviews yielded transcripts as well as responses indexed in Qualtrics. Transcripts were coded at a high level to ascertain topics of importance and novelty (i.e., not identified in the Phase I literature review). Additionally, certain questions in the Qualtrics interview protocol yielded ranking results (e.g., "Rank all of your selected barriers from most to least challenging"). Therefore, we analyzed the ranking results by first considering frequency of selection and then considering their ordinal component. Finally, we supplemented the ranking data analysis with coded transcript data to provide qualitative context.

Results

These interviews allowed us to investigate several core research questions: We examined the alignment between the findings of our Phase 1 literature review and the perspectives of the subjects under investigation. Secondly, we aimed to identify areas of agreement and disagreement between two subsets of our interview subjects: data center experts and owner/operators. Lastly, we sought to characterize the variations in barriers to achieving energy efficiency among different types of data centers, taking into account factors such as DC type, geography, and public/private ownership. The structure of our interviews also informed our findings, relating primarily to procurement, performance metrics, and barriers to energy efficiency. We've grouped these findings according to these sections of the interview protocol. Throughout this section, we will provide context and discussion around the findings from our interviews, drawing connections to the existing literature where appropriate.

Procurement process

Our discussions of the procurement process addressed subject views and experiences related to the specification of equipment, the financing of equipment, and the contracting process. In addition, we queried any procurement tools they used, built, or desired. We separated IT hardware from facility infrastructure and asked our respondents familiar with both to provide detail on the procurement dynamics for each. One major theme to emerge from our interviews was the centrality of vendors to the procurement process for both IT and non-IT hardware. We are emphasizing these findings because multiple interview subjects discussed the influence of vendors without specific priming or prompting by our interview protocol. We designed this protocol to reflect our previous review of the literature. Within that review we identified several procurement-related topics with robust discussion in the body of scholarship, such as the split incentive between facilities and IT operations. We did not include any vendor-related questions or prompts within our interview protocol, but it emerged as a theme as multiple respondents invoked vendors as key drivers of procurement decision making. These respondents included multiple classes of operators as well as experts who themselves provide service/solutions to operators or were familiar with the industry. For these reasons, the breadth of responses on this topic, its emergence without specific priming by our protocol, and the related lack of discussion of vendor relationships in the existing data center literature, we have chosen to emphasize this result.

Vendor interactions drive procurement behavior

The significance of vendor dynamics in the procurement process has emerged as a prominent theme in our findings, despite not being explicitly incorporated into our interview protocol. Through the repeated invocation of this topic by our interview subjects, it became evident that the interactions and relationships with vendors play a crucial role in shaping the procurement landscape. This organic emergence of the vendor dynamics theme underscores its inherent importance and highlights the need for a deeper understanding of its influence on the overall procurement process. Our data center operating subjects highlighted vendors as the key actors in the procurement process. According to one enterprise DC operator, once a need for a purchase was identified internally, existing vendor relationships specified both IT and non-IT purchases:

> *"We set a strategy along a product line and stick with it. The manufacturer may grow and change their products but we'll stick with those compatible elements inside of a product line. We do that with HVAC, we do it with UPS, we do it with server infrastructure, we do it with storage, we even do it with racks and power distribution units."* (O/O 2).

Taken together, the body of interviews revealed that the reliability, interoperability, and familiarity of the largest vendors is a key asset to their customer base. The risks and rewards of marginal improvements are asymmetric in the industry; the downsides of failure far exceed the upsides of marginally more efficient products. In the words of a colocation center operator:

> *"We only work with the most common manufacturers, the biggest manufacturers…We're not really inclined to try some newfangled technology that saves us 2% on power at the cost of massive*

> *outages… [we have] only half a dozen vendors on any piece of equipment and usually fewer. We really don't go outside of those."* (O/O 4)

It is worth noting that our interview subjects were typically purchasing off-the-shelf, commodity style IT hardware products. While this pattern of procurement is overall commonplace within the DC industry as a whole data centers at the technical frontier often purchase custom processors, graphical processing units (GPU) or application-specific integrated circuits (ASIC) they design internally. Investigating the procurement dynamics and the role of vendors in those more customized purchases presents an opportunity for future research to extend or contextualize these findings.

Vendor choice is durable

Given their central role in product specification, vendor selection is the major determinant of data center performance. Rather than make this selection case-to-case, or varying significantly purchase to purchase, our respondents characterized their selection process as centralized and their selections as persistent. They tended to stick with known vendors and product lines until there was a significant failure. According to a colocation operator:

> *"We have relationships that we've had for a long period…Over time, as we've had issues with certain manufacturers, we cross them off the list."* (O/O 4)

The duration of this vendor selection persists year to year, and its length is especially noteworthy given the relatively short lifecycle of IT hardware. Describing their organization's purchasing, an enterprise operator stated:

> *"For many years, our standard server infrastructure was [a well-known IT manufacturer]. At some point that relationship soured and it switched to [another prominent provider]. It switches every 10 or 12 years."* (O/O 2)

The duration of vendor choice in the case of this operator spans multiple generations of product life cycles. This further increases the significance of vendor selection of DC performance. This perspective was echoed by vendors themselves. Commenting on the stickiness of vendor choice, one data center management software provider stated:

> *"You don't want to screw up somebody's data center because it's so critical and a backbone for so many industries…reliability is so important…even if a machine isn't as optimal if you can trust the salesperson, you like them more, you'll buy from that salesperson even if it's more expensive, even if it's not as good."* (E/R 5)

The risks of an outage to an individual operator exceed the rewards of efficiency. Trust that purchasing information is reliable is thus paramount. Accordingly, in the infinite game of ongoing data center operations, maintaining a relationship with a trustworthy vendor is more important than maximizing the performance of any single product.

Vendor relationships are personal

A final point of emphasis by our respondents relating to vendors was the personal nature of these relationships. Individual points of contact at supplier organizations were highlighted as the key vector of interaction. Rather than trust in a given institution, our respondents characterized purchases being made between familiar individuals. In the words of one enterprise operator:

> *"The last time we went through the [specification] exercise to determine small scale UPS technology or racks, in each instance we went to [our main vendor] first and some trusted partner there will kind of help guide us through the industry state of things."* (O/O 2)

The interpersonal trust necessary for buyers to navigate the industry cultivates a rapport between individuals. This was even more strongly emphasized by our subjects who themselves currently were or had served as vendors to the DC industry. They went so far as to say that personal relationship determines purchases on the margin, with one data center management software provider stating:

> *"We had two operators say that it basically just came down to whoever sales person they liked the most was who they bought from. Data cent-*

ers are all relationship based… Now they're not going to go out of the way and buy a piece of [junk] product, if it's all things equal they're going to the salesperson personality [they prefer]; relationship is a big factor." (E/R 5)

The importance of vendors to purchasing is on some level intuitive; the essence of a purchase is a buyer and a seller. Data center hardware is a highly competitive industry, with differentiated products and a short product lifespan. Given the importance of DC services to many organizations, and given the significant expense required to build and operate these facilities, vendors represent a key source of information in a rapidly evolving industry. Vendors are the key determinant of product specification, vendor choice is durable, and vendor selection is built on personal relationships.

Government procurement dynamics are unique

In the course of our research, we spoke with multiple operators of U.S. federal government data centers. These respondents emphasized the influence procurement process regulations had on their purchasing behavior. These responses agree with the robust body of literature describing the influence of the public procurement process and regulatory constraints on purchasing behavior (Payne & Weber, 2012; Telgen et al., 2012; Thomasson and Anna 2016). We're highlighting these results as multiple federal operators described these dynamics influencing the operation of their data center in a manner in broad agreement with the existing scholarship.

Regulation creates threshold costs which substantially shape buying behavior. In one case, a federal scientific computing center operator avoided purchases exceeding a threshold value above which investments could not come out of their general operating fund:

"There's a big incentive for us to generally do projects that can stand alone and will be a total cost of under \$500,000. Doing things over that price tag requires a different color of money." (O/O 3)

This investment included both IT equipment and air handling upgrades. Avoiding a more onerous procurement process was a common goal. A different federal DC operator spoke about specifying products to keep them below a contracting threshold:

"*One of our main goals was making sure that a unit that could be purchased was under \$25,000…because they wanted to be able to purchase without [the more formal] procurement [process].*" (O/O 1)

Above that level, purchases would have to go through a formal request-for-proposals and comply with federal purchasing guidelines. These types of purchases carry both significant marginal administrative burden and increased legal risk. A formal procurement process exposes the agency to the formal bidding process, as well as any potential related litigation. At minimum this delays the purchase. In response to this administrative burden and their own level of familiarity with the contracting process, federal data centers operators defer, avoid, or otherwise modify their purchases (Newkirk et al., 2022). Navigating this procurement process is not within the core competencies of facility or IT professionals.

Performance metrics and monitoring

Energy performance metrics are a common topic of academic research relating to data centers. The industry standard, power usage effectiveness (PUE), is an incomplete measure of energy performance that is often deployed in contexts for which it was not designed and in fact is inappropriate (Yuventi & Mehdizadeh, 2013). There is an expectation in the literature that increased performance would be associated with more comprehensive or better aligned metrics (Horner & Azevedo, 2016; Klemick et al., 2019). Scholars will also design their own metrics, some of which incorporate siting factors (Li et al., 2020) and researchers predict industry energy performance improvements once they coalesce around a single improved metric (Guitart, 2017).

Our interview subjects reported a large variety of metrics in use. Respondents were provided an inventory of the four data center optimization initiative (DCOI) performance metrics[1]: energy performance,

[1] This inventory of metrics was used as the baseline dude to broad industry familiarity with each and to allow comparison to existing research into federal Data Center Optimization Initiative (DCOI) compliance.

uptime and reliability, server utilization, and server virtualization. Respondents were provided an example metric for each of the inventory: power usage effectiveness, annual uptime percentage, utilization percentage, and number of virtual machines per server. There was also space to report their own metrics, both alternative measures within the DCOI categories and for their own alternative measures of site performance, as well as significant opportunity for discussion. The selection frequency and mean ranking of these DCOI metrics, in addition to water usage efficiency, comparing experts to operators is shown in Table 1. There was consistent prioritization of energy performance between subject groups, but operators placed greater emphasis on reliability than did experts.

The exhaustive inventory of respondent-provided metrics is listed below in Table 2. A full definition of each of these metrics can be found in the supplementary materials, Appendix 3. For visual clarity, excepting water usage effectiveness, which is listed above in Table 1, we do not display the significance rankings of respondent-provided metrics.

One theme that emerged from these metrics-related discussions was the heterogeneity of metrics between facilities. There was variety in both the directly reported metrics and in the conversations around metrics more generally. The approach to measuring facility performance depends on facility output: different services have different criteria. An operator who managed data centers for a hospital facility emphasized the need for "24-by-365 reliability", while an efficiency expert contrasted expectations of an enterprise vs. edge facility:

> *"[A video streaming company] has their primary data center, but they do a lot of edge computing because they want to reduce latency so that little round thing that spins around as you're watching your movies doesn't come up … if they lost an edge [DC], they could still get out their product as opposed to somebody like [a large regional bank], if they go down they're in trouble."* (E/R 1)

Even within given performance categories, metrics vary facility to facility: one HPC operator described how optimizing for compute time resulted in a lower net energy use by their DC, despite the higher cooling load:

Table 1 Selection counts and mean prioritization of all performance metrics selected by more than one respondent. Prioritization ranged from 1 to 4, with 1 being the most important metric. Note that respondents could provide their own entries, but the only respondent-originating metric to receive multiple selections was water usage, which was selected by three U.S. operators with facilities west of the Rocky Mountains

Metric	Selection count	Average ranking	Selection count	Average ranking
	Experts ($n=7$)		Owners / Operators ($n=9$)	
Energy Performance	7	2.29	9	2.33
Uptime and Reliability	7	2.57	7	1.14
Server Utilization	7	2.71	4	2.5
Server Virtualization	2	4	2	4
Water Usage	0	–	3	3

Table 2 The full inventory of unique respondent-provided data center performance metrics. Save water usage, which was selected three times, no respondent-provided metric was mentioned more than once. Note the breadth of performance characteristics, ranging from environmental performance to compute time to memory utilization

Respondent class	Uniquely provided metrics
Experts ($n=7$)	Managed Services, Infrastructure Capability, Total Cost, Performance (time-to-model solution)
Owners/Operators ($n=9$)	Queue Length, Delta T, Energy Reuse Efficiency, Office Energy Usage, Storage Utilization, Sewer Usage, Capability Metric, Water Usage Effectiveness (WUE)

"*Optimizing science and engineering…if you run the computers hotter, you can get to solution faster…you use fewer hours to come to the solution but the cost per hour for cooling may be going up.*" (O/O 10)

As previously mentioned, we found no convergence around any individual advanced metric by respondents in our sample. Further, our respondents described the significant internal capacity required for data collection and monitoring. These capabilities are not absent from the industry or these facilities, but selection effects and response bias by our subjects means we were speaking with a sample skewed towards the highest-performing operators.[2] Collecting relevant real-time performance data and ensuring the reliability of those data requires a level of technical proficiency not universal to the industry. Additionally, most advanced composite metrics require core output to be straightforwardly connected to data center energy usage. This is differentially feasible based on organizational purpose as one data center expert indicated:

"[An internet commerce company] is a computer dominated enterprise, and they could convert everything in that enterprise to a per transaction basis. Labor, energy, water, you know everything's a transaction. They had this overall energy metric of Btus-per-transaction, and that's the golden standard." (E/R 1)

This may be impeding the coalescence around next-generation metrics predicted by the literature, and speaks to an advantage of PUE: while not ideal, it is straightforward to collect and interpret, and does not depend on organizational mission.

Another theme relating to performance metrics we found was a changing understanding of reliability among distributed DC service providers. Our subjects highlighted that distributed computing enabled *network-level resiliency* without individual *facility redundancy*. Historically, data centers have achieved reliability through on-site backup power and related infrastructure. This enables facilities to continue to provide services even during electrical system disruptions. Distributing IT demand across multiple networked facilities enabled reliable service without requiring that level of redundant infrastructure at any individual facility, as highlighted by one of our expert respondents:

"All the cloud providers, [the major internet technology companies] they have their redundancy in the network rather than in the data center. They build data centers sometimes without even diesel backup…They're using their network for reliability." (E/R 1)

This characterization was then corroborated by a hyperscale enterprise cloud operator:

"We have an evolving relationship with uptime and reliability. Specifically, moving away from a facility-level view to a network-level view." (O/O 9)

The flexibility afforded by these networked providers enables operators to shift usage according to local grid and climate conditions. This presents an opportunity for operators to engage in demand response, and potentially even optimize siting to respond to real-time energy prices and free-cooling availability, as well as perform other site-specific optimizations.

Barriers to energy efficiency

In the final section of our interviews, we asked our subjects to select the barriers to energy efficiency they had encountered. These barriers were subdivided according to the categories we had identified in the previous phase of this research. Respondents could also include any barrier they had experienced that was not listed in the inventory. After our subjects finalized their selections, we asked them to rank the barriers in descending order of importance. While this task prompted valuable discussion, the small size of our subsamples meant that we could not conduct quantitative analyses on these rankings.

Experts and operators differ in views of barriers

One motivation of this research design was to compare expert and operator perceptions of barriers to energy efficiency. Discrepancies or areas of significant disagreement between these subject groups could indicate gaps in the existing body of knowledge or

[2] To quote one of our operators describing external consultations they perform: "if you walk into a site and you see carpeted floor tiles, you kind of have a sense of what you're dealing with.".

potentially impactful policy interventions. After thorough analysis of our results, it is evident that significant distinction cannot be reported between the views of experts and operators. This conclusion does not stem from uniform responses across the groups but rather from the heterogeneous nature of the responses within the groups. There is one notable exception to this characterization: barriers relating to embedded data centers.

Figure 1 shows a comparison of the quantity of barriers selected as significant by experts and operators. Note that these are the barriers our previous research characterized as relating to baseline resources and capacity. The equivalent comparisons for barriers categorized as Organizational or External context are available in Appendix 1. In only one case did the quantity of selections differ by more than three, in the case of the barrierpL titled *Data Center Embedded in Multi-Use Building*.

What this may indicate is the existence of a mismatch between the mental models of data centers used by efficiency experts and data center practitioners (Morgan et al., 2001). Efficiency experts tend to study facilities on the technical frontier, those with the capital and technical expertise to deploy cutting-edge technologies. The facilities they themselves manage or are consulted on are likely to be large-scale, purpose-built facilities. It simply isn't worth it to hire an expert consultant to optimize the efficiency of a server closet. This presents an opportunity for future research, as formal characterization and comparison between expert and operator mental models is a staple of decision science research (Skarlatidou et al., 2012). This type of mapping could solidify differences in perception and highlight potential opportunities for experts to better serve practitioner needs.

The lack of expert focus on embedded data centers is significant for the simple fact that these facilities have distinct challenges to efficient operation (Dagg et al., 2015). Design optimizations for IT utilization and air-handling best practices are sometimes impractical in embedded facilities. In one memorable case,

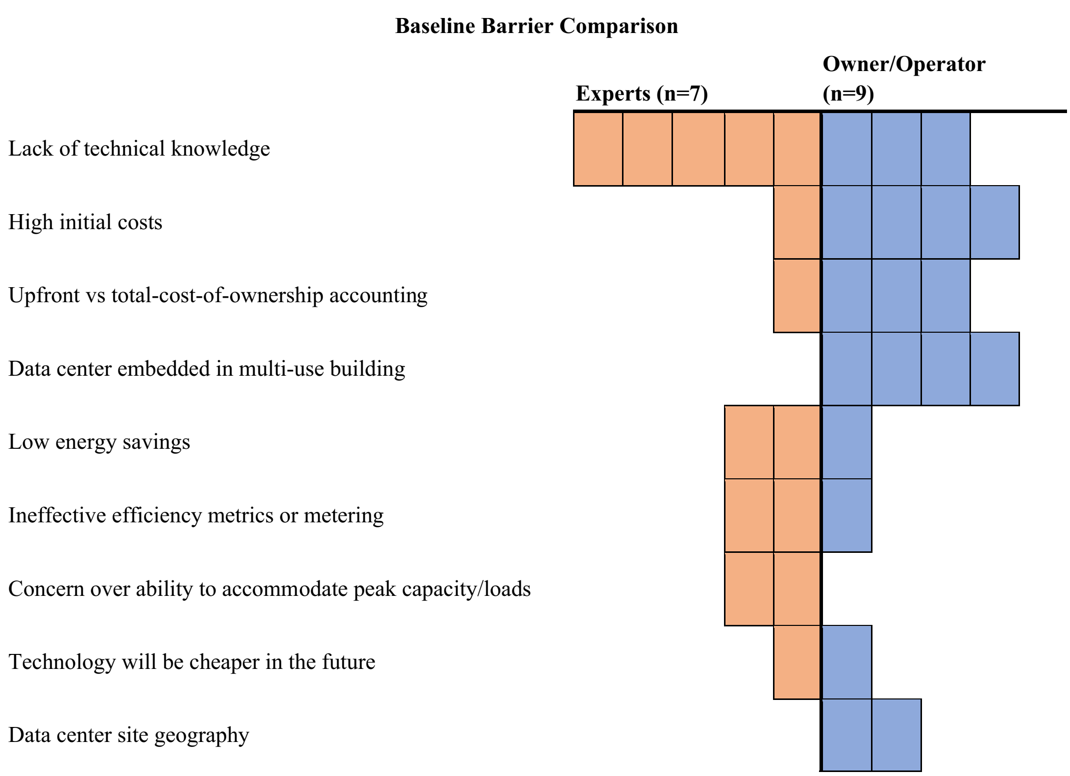


Fig. 1 Comparison of the Baseline and Capacity related barriers selected by experts and operators. Note the barrier selected with greatest frequency by operators, *Data center embedded in multi-use building*, was not selected by experts. Barriers are listed in descending order of total selections, secondary sorted in by quantity of expert selections

one of the embedded centers managed by an interview subject did not have its own distinct chiller zone, severely limiting what efficiency improvements were feasible to implement.

Internal siloing, especially relating to the division between IT and facilities staff, was a consistently reported barrier by both experts and operators. It is then noteworthy that similar vendor dynamics inform purchases of these hardware. What this suggests is these silos impede efficiency more through operational decision making, infrastructure design, and optimization than through equipment purchases.

Barriers differ across data center types

Our interviews also highlighted differences in challenges unique to different types of DC facilities, such as colocation facilities or government data centers. This subsection will include analysis of the distinct challenges encountered in three specific types of data center facilities: colocation facilities, government data centers, and technical frontier hyperscale centers. Our results relating to colocation facilities are broadly consistent with existing scholarship, indicating the continued relevance of these dynamics. We are highlighting the talent attrition and brain drain challenges, which are especially acute in federal data centers, because while that is a known challenge in the federal workforce with an accompanying body of academic literature, it does not typically focus on the IT workforce. We emphasize our hyperscale facility results because they represent a novel phenomenon of compute convergence at hyperscale facilities with only minor and recent discussion in the literature. This convergence between the computational problems of frontier scientific and enterprise computing is driving a convergence on the design of the facilities, a notable shift from industry at the dawn of the cloud era where the largest enterprise centers diverged significantly from HPCs in their heat density and physical footprint. The rise of large language AI models and their accompanying massive parallelized computational training load should only accelerate this trend.

Colocation presents unique challenges

Barriers specific to colocation facilities are a common topic of discussion in the research literature. Such barriers include the split incentive for energy efficiency between facility operators and colocation customers, and pricing structure (Delforge & Whitney, 2014; Klemick et al., 2019). Our interview protocol introduced the topic of barriers related to the colocation business model, and respondents would typically discuss it in passing. We spoke to one dedicated colocation operator as well as an expert who had extensive experience in the industry, including current service contracts. These respondents spoke extensively on colocation topics. Our interview subjects reported colocation-specific barriers to efficiency, including split incentives, with one colocation operator stating:

> *“Why would we, [a Colo] go spend $10,000 on floor tiles if we’re never going to see it back, aside from maybe [making] the customer slightly happier?”* (O/O 4)

Colocation operators also highlighted their legal risk exposure from their contract structure as impeding efficiency. Typical colocation contracts mandate an intake air temperature and humidity level which don’t necessarily reflect ASHRAE best practices, or even the specifications of the IT equipment. If a colocation facility has multiple customer tenants, these contractual obligations aren’t necessarily consistent, which in effect means the facility must cool to the lowest contractual level. Even if it would no effect on performance, deviation from these contractual mandates exposes the operator to potential legal consequences, as highlighted by one of our colocation interview subjects:

> *“All of our customers say they want to be more energy efficient, but they all have contractual [service level agreement] (SLA) that they can sue us for not meeting.”* (O/O 6)

One of the colocation operators spoke about an experience trying to adjust the parameters of an existing contract to match manufacturer guidelines and raise intake temperature. The customer learned of this, threatened legal reprisal if there was any deviation from guidelines, and both the contract and operations remained unchanged.

Government data centers face distinct challenges

Data centers operated by the federal government faced distinct challenges, including that of knowledge loss from human capital attrition, or 'brain drain'. Challenges related to succession planning extend beyond data centers and apply to the federal workforce more broadly. While there is not a robust federal infrastructure for retirement-related knowledge transfer, succession planning efforts are a known correlate of high-performing organizations within the federal government (Goldring, 2015). We spoke with two government data center operators as well as two government-affiliated data center experts. Awareness of attrition issues and workforce human capital was highly salient to one of our interview subjects, a federal data center operator planning for their own retirement:

> *"It's brain drain, the folks that are leaving. Succession planning has to be in place…across all areas of the data center we have a lot of people leaving. I'm actually in that category of being old. So I actually am trying, with three or four people right now, to do succession planning for me."* (O/O 8)

The federal workforce tends to be low churn, making succession a rare and impactful event, and while significant hiring infrastructure is in place, succession-related knowledge transfer lacks formal guidance. Additionally, the aging federal IT workforce is approaching a generational transition, making transfer of human capital vital to maintaining reliability and efficiency (Wesemann, 2022). For federal agencies, consolidation and outsourcing of DC services provides one pathway to addressing these succession-related challenges, effectively obsolescing the need for internal operations capacity. In those cases where internal DC operations must be maintained, staff should review future retirements, implement a formal succession planning strategy, and assign an energy efficiency champion, among other best practices (Dresang, 2023; Loomis, 2017; Wilkerson, 2007). While any individual retirement is likely to result in the loss of embodied and tacit knowledge, the cultivation of organizational resilience through a process of continuous improvement can mitigate these losses. Implementation of an energy management system such ISO 50001 Ready can benefit DC operators both by directly improving energy performance (Better Buildings Initiative, 2018) and by developing organizational capacity (Fuchs et al., 2023).

Hyperscale facilities require significant resources

Our interviews with both experts and practitioners would often address the largest and most computationally intensive facilities. Such facilities include scientific computing facilities,[3] the largest enterprise and cloud computing centers, and large language AI training centers. While these were sometimes opaque references, our interviews included two operators of scientific hyperscale facilities, an operator of a hyperscale enterprise facility, and two government hyperscale experts. These discussions often focused on the underlying capacities of these organizations: to construct and operate a data center of this scale necessarily requires significant financial resources and staff capacity. More interesting was the notion that these facilities are converging in operation and construction. According to one hyperscale computing expert:

> *"Some of the computers that [technical frontier enterprises] are building, you know [an internet giant] is building or [an electric vehicle manufacturer] or [an American semiconductor design firm]. I think they're building supercomputers…think they're running single jobs, and I think that they're running into the same dynamics that [HPC operators] run into."* (E/R 4)

Historically, scholarship has emphasized the contrast between these facility types; the different purposes of HPC and large enterprise required different approaches to efficiency (Wilde et al., 2014). While not yet thoroughly discussed, there is some literature that hypothesizes this development as a consequence of computational parallelization. Scientific physical systems modeling typically takes the form of a large number of parallel tensor operations. The innovation of the transformer machine learning architecture effectively enabled more efficient parallelization, enabling larger data sets, more neural layers, and orders of magnitude more parameters. The vector operations of these transformer models are analogous to the tensor products in physical modeling, and are performed on the same hardware stack of GPUs and ASICs. This pattern holds across a variety of frontier computational applications: advanced graphical processing or ultra-high-resolution seismographic analysis are each analogous to physical

[3] Sometimes referred to as high-performance computing (HPC) facilities.

systems modeling in their underlying computation (Govind et al., 2023; Krueger et al., 2011). One hyperscale operator emphasized the heat density of parallel processing hardware as the key factor driving facility convergence:

> *"These new technologies, whether it's quantum computing or AI or ML, these things are expanding and growing. You will see the facilities have the same problems: if you're building something for AI and ML, you're probably GPU. And if you're using GPUs, they're really heavy heat…How do you get rid of the heat? It's not surprising that these things are the same problems [in HPC]… the question really is heat density and power density."* (O/O 10)

There is some brief discussion of this software and problem side convergence in recent literature (Hoefler et al., 2022). As cutting edge enterprise facilities devote themselves to parallelized problems, those facilities will come to resemble legacy scientific computing centers in their hardware, physical footprint, and air (Ozawa et al., 2019). This then makes more pressing the issue of disciplinary siloing within the data center industry. Silos exist between facility types, and operational knowledge is closely guarded within the industry. As stated by one expert respondent:

> *"[Propriety around IP impedes best practice sharing]. I'd really like to know what's going on [at the major cloud computing and AI firms], they're not interested, they don't want to share, that's one of their value-adds."* (E/R 4)

While not unreasonable, this jealous guarding of operational knowledge directly impedes industry-level efficiency. As the largest and most technologically sophisticated facilities converge in the types of problems they compute, channels to transfer knowledge and best practices between the largest enterprise and scientific facilities increase in importance.

Siting is a common barrier

Another key insight of our interviews was in exploring data center site selection. There is a robust research literature linking site characteristics to DC energy (Jones, 2018; Turek & Radgen, 2021), water (Chen & Wemhoff, 2022; Siddik et al., 2023), and carbon (Gao et al., 2012; Li et al., 2020; Yu et al., 2023) performance. What factors determine DC sites, the organizational process of siting, and what constrains site selection is a gap in the existing literature. This gap is why we're highlighting our siting-related results. Robust scholarship has characterized site selection as one of the most significant determinants of data center energy and water performance. What this literature does not discuss is the practical determinants of data center siting.

Our subjects characterized the site selection process, and provided insight into what factors restricted their choices. Legacy building stock and customer proximity were common siting factors, especially in organizations where the data center is not central to the core mission. According to one of our enterprise respondents, "that center was born in the closet" and that legacy carries a certain inertia, as well as many organizations not having the capital flexibility to consider site diversification. According to another enterprise provider, colocation and enterprise IT firms also have strong preferences for customer proximity:

> *"I think the majority of data centers are actually being put in the wrong place: We see data centers trying to basically crowd into the middle of metros which makes no sense at all… realistically data centers should be on the periphery of cities and tied directly to the high power grids… Pre-existing facilities especially have no incentive to leave the market [and their current] location."* (O/O 7)

Owing to their economies of scale, access to capital, and high levels of existing expertise, hyperscale operators are able to overcome some of these constraints. HPC simulations aren't latency sensitive applications, so they are able to locate where electric grids and climatological conditions are favorable. In the case of enterprise hyperscale/cloud providers, they benefit from the flexibility afforded a remote service provider to consider those factors in site selection. While hyperscale centers are able to avoid some of these siting constraints, they still are subject to a regulatory environment which can lead to sub-optimal siting, in the words of one enterprise hyperscale operator:

> *"[A barrier to efficiency is] if you are required to have proximity to a certain region for XYZ reason, and the climate or the conditions on*

that region leads to a lower efficiency for the center…[Let's say for instance] you're contracted to be the flagship HPC facility of Spain, you're not actually allowed to site it outside of Spain. You have multiple countries that will not allow you to relocate [even distributed services] outside of the country." (O/O 9)

The regulatory environment that in part determines DC site selection is multifaceted, ranging from local land use and zoning codes to HIPAA compliance and requirements of state sports gambling statutes mandating data be physically stored within particular state boundaries.

Discussion

These interviews provide some suggestions for policymakers as well as areas for future research. This section will overview the limitations of our methods and data, followed by a discussion of potential policy implications, and conclude with areas for future study.

Limitations

The most straightforward limitation to our research design is simple lack of scale. While we attempted wide initial outreach, overall response levels were low, even within snowball sample contacts. While we were able to collect rich and detailed data from each respondent, with interviews lasting on the order of an hour, the overall quantity of subjects remains a modest 16. This limited quantity of respondents should preclude drawing firm quantitative conclusions about the behaviors of different subsets of our sample.

In constructing our sample, we sought to engage with subjects who reflected the breadth of the industry, as well as the breadth of expertise ranging from academic experts to design engineer and efficiency consultants. While we were able to capture a diverse set of practitioners, this methodology is necessarily subject to response bias. Given the significant effort required on the part of our subjects, they are likely to be disproportionately interested in data center efficiency and knowledgeable about operations. This respondent characteristic should also cause us to select for the best managed, highest human capital facilities and organizations. Additionally, while our sample was diverse, it was not representatively weighted by data center type. These sample characteristics should be considered when assessing these facilities or identifying industry trends.

Policy implications

This research revealed information relevant to both the design and implementation of policy. One such area relates to vendors and procurement. Given the central role they play in the procurement process, vendors could serve as key partners in policy implementation efforts. Identified vendors of efficiency products could be key partners in the implementation industry trainings and workforce development programs such as the Data Center Energy Practitioner (DCEP) Program (Radhakrishnan, 2012). This would serve to both directly network efficient suppliers to customers, as well better propagate information about efficient operation and design. Indeed, vendors as effective vectors of efficiency-related information was one of the key findings of the scholarship on energy efficiency market transformation (Blumstein et al., 2000; Duke & Kammen, 1999; Geller & Nadel, 1994).

In addition to equipment manufacturers and distributors, Energy Service Companies (ESCO) could serve as partners in efficiency-related implementation and outreach. ESCOs provide efficiency-as-a-service through energy savings performance contracts, where they finance and implement efficiency retrofits. Data centers have been previously identified as one of the main opportunities for growth in the ESCO industry, and could serve as demand drivers for efficient cooling and air handling technologies (Larsen et al., 2017). Policymakers identifying vendors of efficient products would also enable more intuitive and streamlined procurement tools; the high administrative burden of efficient product specification and selection has been identified previously as a barrier to efficiency (DeMates & Scodel, 2017).

Additionally, guidance to data center operators should include discussion of site selection resources. Siting is fixed for operators of legacy facilities, but any greenfield development or expansion of existing facilities should explicitly consider the implications of siting. This guidance would require corresponding tools to model the impacts of a given site, but policy should reflect the outsize impact of site selection on environmental performance. Different facilities have different siting constraints and possesses different environmental impacts. The latency sensitivity of the actual IT load within a data center is often a key consideration. Large-scale data centers are much more efficient in practice than the equivalent compute spread across diffuse facilities. This means that consolidation of IT demand in hyperscale facilities does reduce environmental footprint, but that also makes the site selection of those facilities extremely influential to overall carbon and water impact based on the resource intensity of available grid electricity mix. Dedicated AI training data centers are hyperscale facilities uniquely suited to environmentally optimal site selection. The heat density of specialized AI training chips makes them ideal matches for the highest efficiency, liquid cooling systems (Zhang et al., 2022). AI training runs can last weeks or months and require significant postprocessing. This makes these workloads uniquely latency insensitive, allowing siting based on free-cooling, and the carbon and water intensity of the electric grid. The minimal marginal impact of a delay of hours in a training run lasting months also means that these facilities are a strong candidate for demand response and carbon aware scheduling (Chien et al., 2023). In contrast, scientific computing facility sites are often chosen based on political considerations. Characterizing performance sensitivities by DC type could enable designers to better understand the siting options for their specific data center.

While these interviews often emphasized non-technical barriers to energy efficiency, that is not to say technical skills to optimize design, operation, and equipment selection are not core to data center performance. Indeed, this work emphasized the siloing between relevant technical domains within organizations as a key impediment to DC efficiency. Policymakers could develop operational best practices documentation that emphasizes the interdisciplinary expertise required for efficient facilities. Finally, policymakers should extend existing succession planning resources to the public sector IT workforce. While not exhaustive, these represent the most directly actionable policy insights of this research.

Future work

Our interview findings relating to network resiliency present several pathways for future study. One relates to the actual design of such data centers: what sorts of infrastructure, what sorts of compute, and what operational changes are necessary to achieve reliable service this way? What are barriers to transitioning from facility redundancy to network resiliency, and how might they be overcome? What are design best practices of the individual facilities and the computing service network as a whole? Additionally, there are theoretical non-energy benefits associated with network resilience; research should characterize these benefits to allow operators to better scope the costs and benefits of this approach. Finally, scholarship should characterize organizations which have adopted this model to identify factors which enable transition away from facility redundancy.

The convergence of heat-dense parallelized compute facilities discussed in these interviews suggests another avenue for future inquiry. There is explosive growth in these facilities resulting from investment in AI, including by those without previous experience designing or operating these types of data centers (Loten, 2023). Researchers should characterize the organizational dynamics and determinants of performance within these data center operators. Additionally, this work has highlighted the key influence of vendor relationships on facility performance. Massively parallelized facilities depend on distinct computational hardware from traditional enterprise centers. The GPUs, ASICs, and AI accelerators in these facilities are supplied by only a few manufacturers, and are often designed in-house by the end use firm. The heat density of these processors also makes them conducive to liquid cooling and heat recovery infrastructure, further distinguishing their suppliers from conventional enterprise facilities (Santosh Janardhan, 2023). Future study should investigate the procurement and vendor dynamics of these facilities and compare them to those identified in these interviews. While not an exhaustive inventory of all potential follow-on research, these represent two pathways for valuable scholarship building on these results.

Conclusion

Data centers consume 2–5% of US electricity, with increasing demand from computationally intensive services such as artificial intelligence. We previously conducted a literature review and identified barriers to efficiency in existing scholarship. We then conducted a series of interviews to compare these findings with the views and experiences of both data center experts and operators. We contacted our interview subjects within our personal and professional network, and then used snowball sampling for further recruitment. Through this process, we conducted formal interviews with seven experts and nine operators. These included interviewers with operators of enterprise, colocation, scientific, and hyperscale cloud computing facilities.

Our interviews were divided into sections on procurement, performance metrics, and barriers to energy efficiency. We identified vendors as the key actors in procurement. In contrast to expectations in the scholarly literature, we did not observe a convergence around advanced efficiency metrics—instead, advanced metrics were facility- and application-specific. We did find convergence in the design of facilities for leading edge computation across organizational types due to the applicability of heat-dense, parallelized hardware to these computational loads.

While our interviews provided us with rich insights from our subjects, those subjects were not a representative sample of either the public at large or the data center industry. Our respondents operate disproportionately efficient facilities and are disproportionately interested in data center energy use. Despite these limitations in our sample, we were able to identify relevant trends in the industry and subjects of disagreement between the literature and our interview subjects. Through this research we identified opportunities for both policy interventions to drive efficiency as well as future research. Guidance for data center facility design and operation should include site selection analysis; siting is the single most significant factor determining the carbon and water footprint of data centers. Future study should examine the potential for demand response in networked service providers, and vendor dynamics in selection of cooling technologies for the heat-dense parallel applications that will account for significant medium-term demand growth.

Acknowledgements This manuscript has been authored by an author at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges, that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

The authors would like to thank Steve Greenberg, Dale Sartor, Sara Smith, Arman Shehabi, and Heidi Fuchs of the Lawrence Berkeley National Lab for their guidance and support, as well as all staff members at the Center

of Expertise for Energy Efficiency in Data Centers.

Author contribution All authors contributed to research design. Chris Payne acquired the funding. Alex Newkirk conducted the interviews and processed the interview data. All authors contributed to data analysis. Alex Newkirk and Nichole Hanus drafted the manuscript. All authors reviewed and commented on manuscript drafts.

Funding Open Access funding provided by Carnegie Mellon University. This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Federal Energy Management Program, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

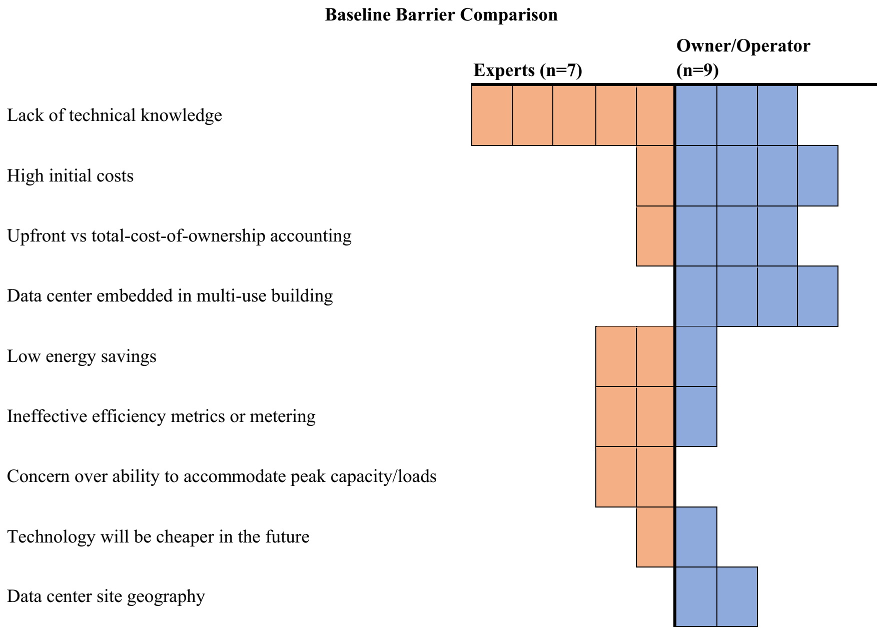
Data availability The library of sources used in this research are publicly available in a Zotero library, accessible at https://www.zotero.org/groups/5384469/overcoming_social_and_organizational_barriers_to_data_center_energy_efficiency.

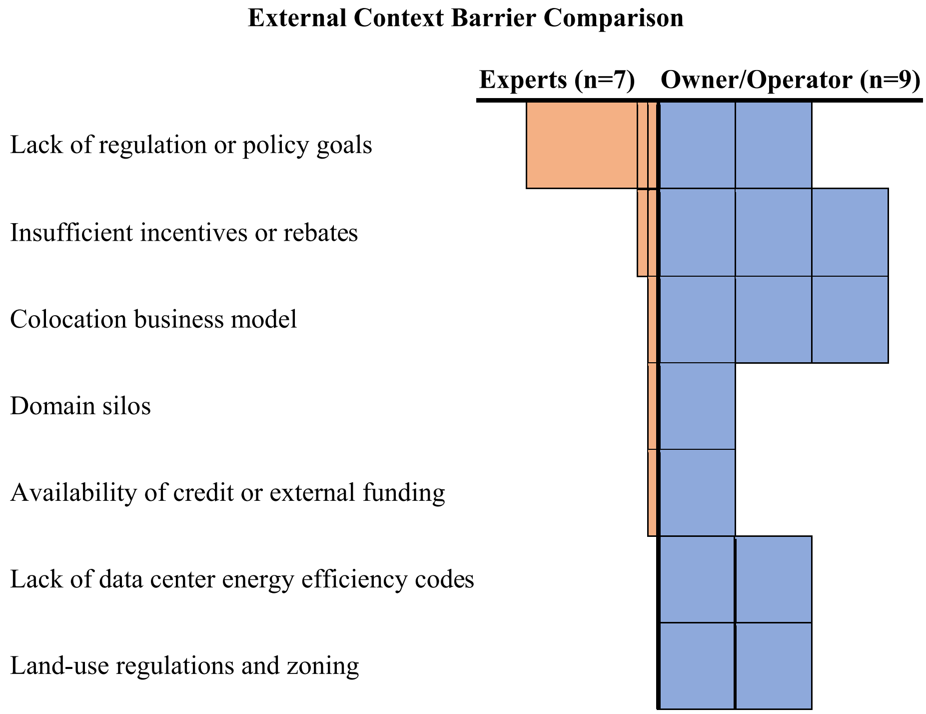
Declarations

Competing interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Appendix 1 Organizational and external context barrier comparison

Below are figures comparing the expert and operator selections of barriers to energy efficiency in data centers classified in our previous work as *Organizational* or *External Context* related.





Appendix 2: Respondent-Characteristics

Respondent ID	Interview order	Age range	Education level	Professional role	Procurement domain	Gender	Organization type
Expert 1	1	65 and above	Graduate or Professional Degree	Staff Scientist	Facilities	M	Public Sector Laboratory
Enterprise 1	2	35–44	Graduate or Professional Degree	Chief Information Officer	Both IT and Facilities	M	Healthcare Network
Hyperscale 1	3	45–54	Bachelor's Degree	Head of IT Department	Both IT and Facilities	M	Public Sector Laboratory
Colocation 2	4	35–44	Bachelor's Degree	Critical Facilities Engineer	Facilities	M	Colocation Provider
Expert 2	5	35–44	Graduate or Professional Degree	CEO	Facilities	F	Energy Startup
Expert 3	6	55–64	Bachelor's Degree	Director of Data Center Design	Both IT and Facilities	M	DC Services Consultancy
Hyperscale 2	7	45–54	Graduate or Professional Degree	HPC Researcher	Both IT and Facilities	M	Public Sector Laboratory
Colocation 1	8	35–44	Bachelor's Degree	Sales Engineer	Facilities	M	Colocation Provider
Expert 4	9	55–64	Bachelor's Degree	Co-Chair	Both IT and Facilities	F	DC Energy Efficiency Advocacy Org

Respondent ID	Interview order	Age range	Education level	Professional role	Procurement domain	Gender	Organization type
Enterprise 2	10	35–44	Some College	CEO & President	Both IT and Facilities	M	Enterprise DC Provider
Hyperscale 3	11	55–64	Bachelor's Degree	Chief HPC Engineer	Both IT and Facilities	F	Public Sector Laboratory
Expert 5	12	25–34	Graduate or Professional Degree	CTO and Head of AI	Both IT and Facilities	M	DC Monitoring Software Developer
Expert 6	13	25–34	Bachelor's Degree	Customer Success Engineer	IT	M	SaaS Infrastructure provider
Hyperscale 4	14	35–44	Bachelor's Degree	Data Center Staff Engineer	IT	M	Hyperscale Cloud Provider
Hyperscale 5	15	65 and above	Graduate or Professional Degree	Head of Computing Capability	Both IT and Facilities	M	Scientific Computing Center

Appendix 3: Respondent-Provided metric definitions

Find below the definition of each respondent-provided metric.

Respondent metric	Definition
Managed Services	Count of cloud services hosted on some particular DC infrastructure
Infrastructure Capability	Refers to the ability of the data center infrastructure to support the IT equipment and workloads
Total Cost	Total cost of ownership for the data center
Queue Length	Total jobs in scientific computing queue
Performance (time-to-model solution)	Average time to solution for computational physical models
Delta t	Difference in temperature between the inlet and outlet air of an IT equipment
Office Energy Usage	Amount of energy used by the office space in a data center. It includes the energy used by lighting, HVAC, and other office equipment
Energy Reuse Efficiency	Quantity of the energy used in a data center is reused. For example, waste heat from cooling solutions can be reused to heat offices or water
Storage Utilization	Percentage of computer memory being used in a data center
Sewer Usage	Amount of water data center is discharging into the sewer system
Capability Metric	Future-facing metric that scales the amount of IT workload which the networking and HVAC infrastructure can support
Water Usage Effectiveness (WUE)	Amount of water used by a data center per unit of IT equipment power

References

Berg, B. L., & Lune, H. (2011). *Qualitative Research Methods for the Social Sciences* (8th ed.). Pearson.

Better Buildings Initiative. (2018, October 24). *Google Uses ISO 50001 at Global Data Centers for Major Savings |Better Buildings Initiative*. https://betterbuildingssolutioncenter.energy.gov/iso-50001/resources/googleiso-50001-case-study. Accessed 24 Oct 2018

Blumstein, C., Goldstone, S., & Lutzenhiser, L. (2000). A theory-based approach to market transformation. *Energy Policy, 28*(2), 137–144. https://doi.org/10.1016/S0301-4215(99)00093-2

Chen, L., & Wemhoff, A. (2022). *Characterizing Data Center Cooling System Water Stress in the United States*. 2022 ASHRAE Winter Conference.

Chien, A. A., Lin, L., Nguyen, H., Rao, V., Sharma, T., & Wijayawardana, R. (2023). Reducing the Carbon Impact of Generative AI Inference (today and in 2035). *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, 1–7. https://doi.org/10.1145/3604930.3605705

Dagg, J., Lakos, W., & Hammond, S. (2015, May 28). *Managing Data Centers in Multi-Use Environments*. Better Buildings Summit. https://betterbuildingssolutioncenter.energy.gov/sites/default/files/Thursday%20-%20Managing%20Data%20Centers%20in%20Multi-Use%20Environments.pdf. Accessed 20 June 2023

Delforge, P., & Whitney, J. (2014). *Data center efficiency assessment scaling up energy efficiency across the data center industry: evaluating key drivers and barriers*. Natural Resource Defence Council.

DeMates, L., & Scodel, A. (2017). *From Policy to Compliance: Federal Energy Efficient Product Procurement* (LBNL--1003934, 1378568; p. LBNL--1003934, 1378568). https://doi.org/10.2172/1378568

Dresang, D. L. (2023). Succession Planning. In Personnel Management in Government Agencies and Nonprofit Organizations (6th ed., pp. 121–143). Routledge.

Duke, R., & Kammen, D. M. (1999). The economics of energy market transformation programs. *Energy J, 20*(4), 15–64. https://doi.org/10.5547/ISSN0195-6574-EJ-Vol20-No4-2

Fuchs, H., Therkelsen, P., Miller, W. C., Siciliano, G., & Sheaffer, P. (2023). ISO 50001-based energy management systems as a practical path for decarbonization: initial findings from a survey of technical assistance cohort participants. *Energies, 16*(14), 5441. https://doi.org/10.3390/en16145441

Gao, P. X., Curtis, A. R., Wong, B., & Keshav, S. (2012). It's not easy being green. *ACM SIGCOMM Comput Commun Rev, 42*(4), 211–222. https://doi.org/10.1145/2377677.2377719

Geller, H., & Nadel, S. (1994). Market transformation strategies to promote end-use efficiency. *Ann Rev Energy Environ, 19*(1), 301–346. https://doi.org/10.1146/annurev.eg.19.110194.001505

Goldring, C. C. (2015). *A Design for Federal Government Leaders: Succession Planning Through Knowledge Management* [Ed.D., Wilmington University (Delaware)]. https://www.proquest.com/docview/1688688202/abstract/DE23F5CB72A44892PQ/1. Accessed 22 June 2023

Govind, A., Bhalachandra, S., Zhao, Z., Rrapaj, E., Austin, B., & Nam, H. A. (2023). Comparing Power Signatures of HPC Workloads: Machine Learning vs Simulation. *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, 1890–1893. https://doi.org/10.1145/3624062.3624274

Guitart, J. (2017). Toward sustainable data centers: a comprehensive energy management strategy. *Computing, 99*(6), 597–615. https://doi.org/10.1007/s00607-016-0501-1

Hanus, N., Newkirk, A., & Stratton, H. (2023). Organizational and psychological measures for data center energy efficiency: Barriers and mitigation strategies. *Energy Efficiency, 16*(1), 1. https://doi.org/10.1007/s12053-022-10078-1

Hanus, N., Wong-Parodi, G., Small, M. J., & Grossmann, I. (2018). The role of psychology and social influences in energy efficiency adoption. *Energy Efficiency, 11*(2), 371–391. https://doi.org/10.1007/s12053-017-9568-6

Hoefler, T., Hendel, A., & Roweth, D. (2022). The convergence of hyperscale data center and high-performance computing networks. *Computer, 55*(7), 29–37. https://doi.org/10.1109/MC.2022.3158437

Horner, N., & Azevedo, I. (2016). Power usage effectiveness in data centers: Overloaded and underachieving. *The Electricity Journal, 29*(4), 61–69. https://doi.org/10.1016/j.tej.2016.04.011

Santosh Janardhan. (2023). Reimagining Our Infrastructure for the AI Age. *Meta*. https://about.fb.com/news/2023/05/metas-infrastructure-for-ai/. Accessed 28 Aug 2023

Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. *Nature, 561*(7722), 163–166. https://doi.org/10.1038/d41586-018-06610-y

Klemick, H., Kopits, E., & Wolverton, A. (2019). How do data centers make energy efficiency investment decisions? qualitative evidence from focus groups and interviews. *Energy Efficiency, 12*(5), 1359–1377. https://doi.org/10.1007/s12053-019-09782-2

Krueger, J., Donofrio, D., Shalf, J., Mohiyuddin, M., Williams, S., Oliker, L., & Pfreund, F.-J. (2011). Hardware/software co-design for energy-efficient seismic modeling. *Proceedings of 2011 International Conference for High*

Performance Computing*, Networking, Storage and Analysis*, 1–12. https://doi.org/10.1145/2063384.2063482

Larsen, P. H., Carvallo Bodelon, J. P., Goldman, C. A., Murphy, S., & Stuart, E. (2017). *Updated Estimates of the Remaining Market* Potential *of the U.S. ESCO Industry*. https://escholarship.org/uc/item/99033667. Accessed 1 Nov 2023

Li, J., Jurasz, J., Li, H., Tao, W.-Q., Duan, Y., & Yan, J. (2020). A new indicator for a fair comparison on the energy performance of data centers. *Appl Energy, 276*, 115497. https://doi.org/10.1016/j.apenergy.2020.115497

Loomis, S. G. (2017). *Federal Employee Retirement: An Explanatory Case Study on Succession Planning in the Department of the Navy* [D.M., University of Phoenix]. https://www.proquest.com/docview/1925336199/abstract/86D237F081B14C26PQ/1 Accessed 22 June 2023

Loten, A. (2023). AI-Ready Data Centers Are Poised for Fast Growth. *Wall Street Journal*. https://www.wsj.com/articles/ai-ready-data-centers-are-poised-for-fast-growth-fadae952. Accessed 28 Aug 2023

Niall McCarthy. (2021, February 10). *Which Countries Have The Most Data Centers?* [Digital Image]. Statista Daily Data. https://www.statista.com/chart/24149/data-centers-per-country. Accessed 25 Jan 2024

Morgan, M. G., Fischhoff, B., Bostrom, A., & Atman, C. J. (2001). *Risk Communication: A Mental Models Approach* (1st edition). Cambridge University Press.

Newkirk, A., Watson, K., & Robinson, G. (2022, August 21). The Chilling Effects of Administrative Burden on Efficiency Policy Uptake: Examining the Case of Federal ESAs. *Summer Study on Energy Efficiency in Buildings*. ACEEE Summer Study on Energy Efficiency in Buildings. https://escholarship.org/uc/item/2401896c. Accessed 22 June 2023

Ozawa, M., Chaplin, J., Pollitt, M., Reiner, D., & Warde, P. (2019). *In search of good energy policy*. Cambridge University Press.

Payne, C., & Weber, A. (2012). *Public Sector Procurement: Issues in Program Development & Delivery* (LBNL--6015E, 1172252; p. LBNL--6015E, 1172252). https://doi.org/10.2172/1172252

Precedence Research. (2023). *Data Center Solutions Market Size To Hit USD 81.13 Bn By 2032*. https://www.precedenceresearch.com/data-center-solutions-market. Accessed 25 Jan 2024

Radhakrishnan, B. D. (2012). Data Center IT Equipment Energy Assessment Tools: Current State of Commercial Tools, Proposal for a Future Set of Assessment Tools (LBNL--6252E, 1172120; p. LBNL--6252E, 1172120). https://doi.org/10.2172/1172120

Shehabi, A., Smith, S. J., Masanet, E., & Koomey, J. (2018). Data center growth in the United States: Decoupling the demand for services from electricity use. *Environ Res Lett, 13*(12), 124030. https://doi.org/10.1088/1748-9326/aaec9c

Shehabi, A., Smith, S., Sartor, D., Brown, R., Herrlin, M., Koomey, J., Masanet, E., Horner, N., Azevedo, I., & Lintner, W. (2016). *United States Data Center Energy Usage Report* (LBNL--1005775, 1372902; p. LBNL--1005775, 1372902). https://doi.org/10.2172/1372902

Siddik, M. A. B., Amaya, M., & Marston, L. T. (2023). The water and carbon footprint of cryptocurrencies and conventional currencies. *J Clean Prod, 411*, 137268. https://doi.org/10.1016/j.jclepro.2023.137268

Skarlatidou, A., Cheng, T., & Haklay, M. (2012). What do lay people want to know about the disposal of nuclear waste? A mental model approach to the design and development of an online risk communication. *Risk Anal: Off Publ Soc Risk Anal, 32*(9), 1496–1511. https://doi.org/10.1111/j.1539-6924.2011.01773.x

Telgen, J., Harland, C., & Knight, L. (2012). Public procurement in perspective. In Public procurement (2nd ed., pp.44–52). Routledge.

Thomasson, H. K., & Anna. (2016). *Innovation in the public procurement process: A study of the creation of innovation-friendly public procurement*. In Innovation in Public Services.

Turek, D., & Radgen, P. (2021). Optimized data center site selection—Mesoclimatic effects on data center energy consumption and costs. *Energy Efficiency, 14*(3), 33. https://doi.org/10.1007/s12053-021-09947-y

Wesemann, A. (2022). The performance rewards of human capital development in the federal government. *Public Pers Manag, 51*(2), 151–169. https://doi.org/10.1177/00910260211039876

Wilde, T., Auweter, A., & Shoukourian, H. (2014). The 4 Pillar Framework for energy efficient HPC data centers. *Comput Sci - Res Dev, 29*(3), 241–251. https://doi.org/10.1007/s00450-013-0244-6

Wilkerson, B. (2007). Effective succession planning in the public sector (pp. 1–7). Wattson Wyatt Wordlwide. https://4cleanair.org/wpcontent/uploads/Chicago/EffectiveSuccessionPlanningforPublic.pdf. Accessed 2 Feb 2024

Yu, X., Hu, Y., Zhou, D., Wang, Q., Sang, X., & Huang, K. (2023). Carbon emission reduction analysis for cloud computing industry: Can carbon emissions trading and technology innovation help? *Energy Economics*, 106804. https://doi.org/10.1016/j.eneco.2023.106804

Yuventi, J., & Mehdizadeh, R. (2013). A critical analysis of power usage effectiveness and its use in communicating data center energy consumption. *Energy Build, 64*, 90–94. https://doi.org/10.1016/j.enbuild.2013.04.015

Zhang, Y., Zhao, Y., Dai, S., Nie, B., Ma, H., Li, J., Miao, Q., Jin, Y., Tan, L., & Ding, Y. (2022). Cooling technologies for data centres and telecommunication base stations – A comprehensive review. *J Clean Prod, 334*, 130280. https://doi.org/10.1016/j.jclepro.2021.130280