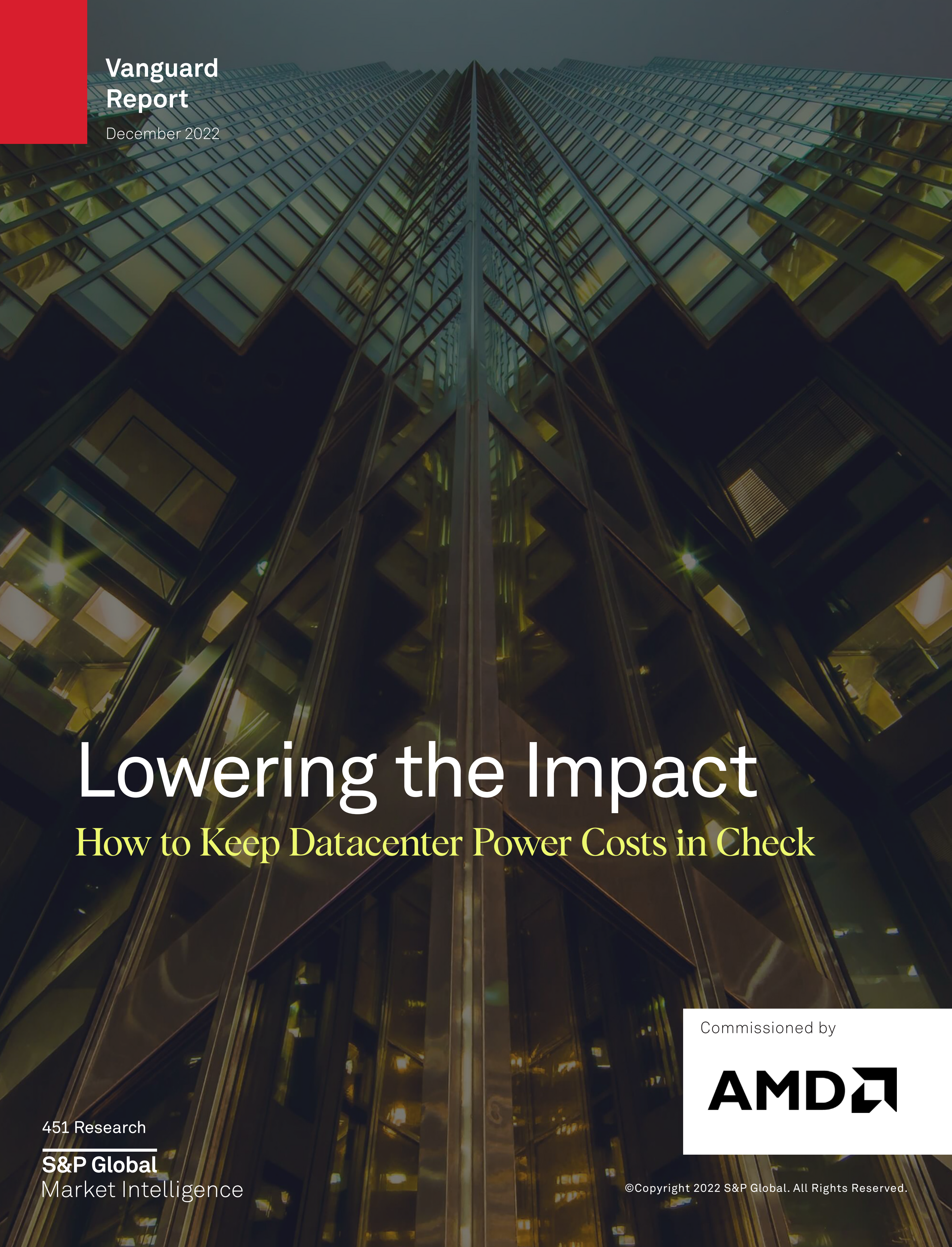


Vanguard Report

December 2022

# Lowering the Impact

## How to Keep Datacenter Power Costs in Check

Commissioned by AMD

451 Research

S&P Global Market Intelligence

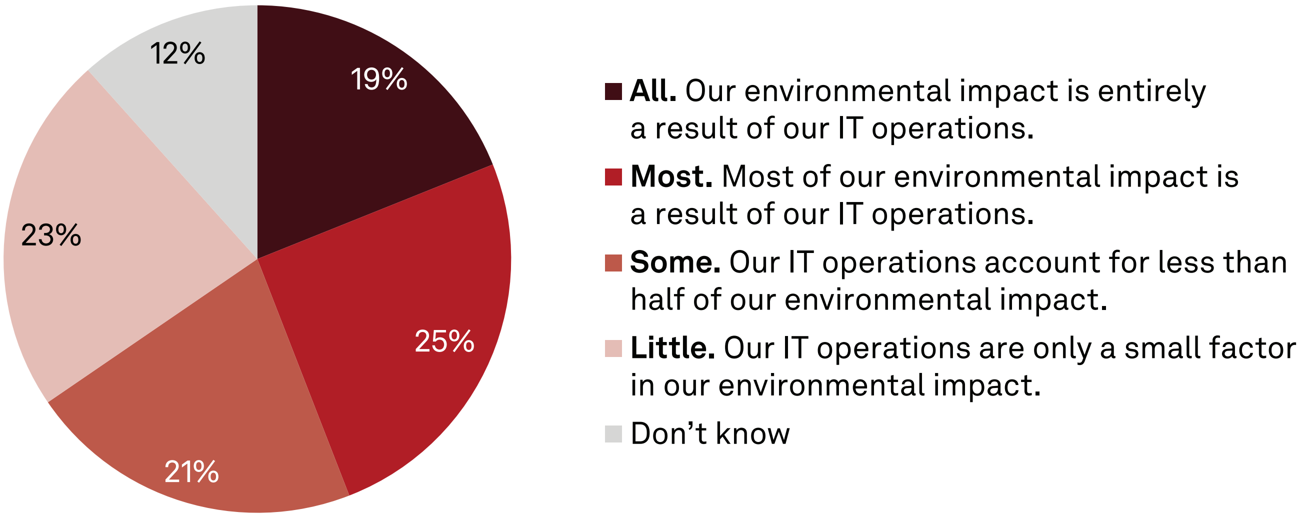
# Introduction

Enterprises across the globe have never been more aware that they must take full responsibility for managing, and reducing, their environmental impact. If they don't, investors could penalize them, governments may step in to regulate, and customers could start to look elsewhere for cleaner, more efficient suppliers. As more companies set climate targets, it is unclear whether the good intentions are enough to reflect the ambition — and the measurable results — required to achieve steep greenhouse gas emission reductions as a matter of considerable urgency.

To improve the situation, one of the first things companies must get to grips with is the efficiency of their IT infrastructure. There is a growing recognition of this fact: In the July 2022 451 Research Digital Pulse user survey, nearly half of the IT decision-makers polled said that IT operations now account for most (25% of respondents) or all (19%) of their environmental impact. Our survey shows that many companies already have formal targets in place for reducing environmental impact, and they consider power efficiency and reduced emissions important factors influencing their technology buying decisions. Many also express a willingness to pay a premium for products that can deliver a reduced impact. And they expect their vendors to provide the necessary tools for tracking that impact.

But the benefits here go beyond appeasing investors and exhibiting good corporate citizenship. In the specific context of IT, efficiency achieved by efforts to reduce environmental footprint will lead directly to meaningful cost savings, while the increasingly global issue of rising energy prices only makes the total-cost-of-ownership gains more pronounced.

**Figure 1: Environmental Impact of IT Operations**



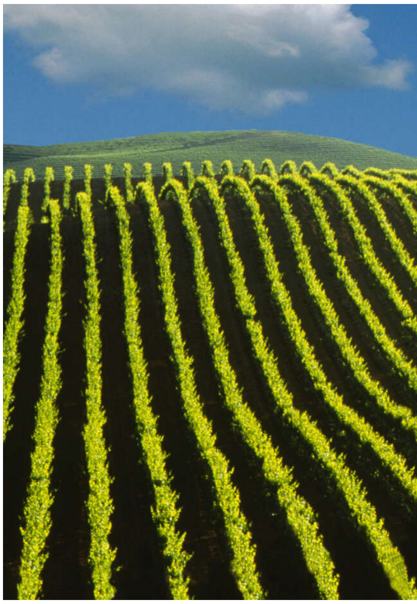

Q. What portion of your organization's total environmental impact would you say is a result of its IT operations?
Base: All respondents (n=576)
Source: 451 Research's Voice of the Enterprise: Digital Pulse, Environmental Impact 2022

# The Take

451 Research estimates that in 2019 all datacenters in the U.S. — including everything from enterprise facilities to server rooms and closets, as well as multi-tenant, colocation-type facilities — consumed approximately 268TWh of energy. That represented 6.3% of total U.S. energy usage. If datacenters and server rooms in the U.S. represented a country, that country would rank above Mexico in terms of total energy consumption. Since 2019, the compound annual growth rate of datacenter energy consumption has been 2.2%. Compare this figure with the total growth in demand for power over the same period — 0.3% overall — and it's clear that the amount of power datacenters consume is growing much faster than the national aggregate average.

Governing bodies have certainly noticed. In some cases, they are even stepping in to put the brakes on growing datacenter power consumption. But environmental initiatives too often amount to lip-service commitments that are little more than "greenwashing." Offsets are not a silver bullet. Real action with measurable results is needed. Datacenter operators and providers must adopt a posture of continuous improvement, using new methods for efficiency gains in addition to decarbonization — in that order. Sustainability must be a priority.

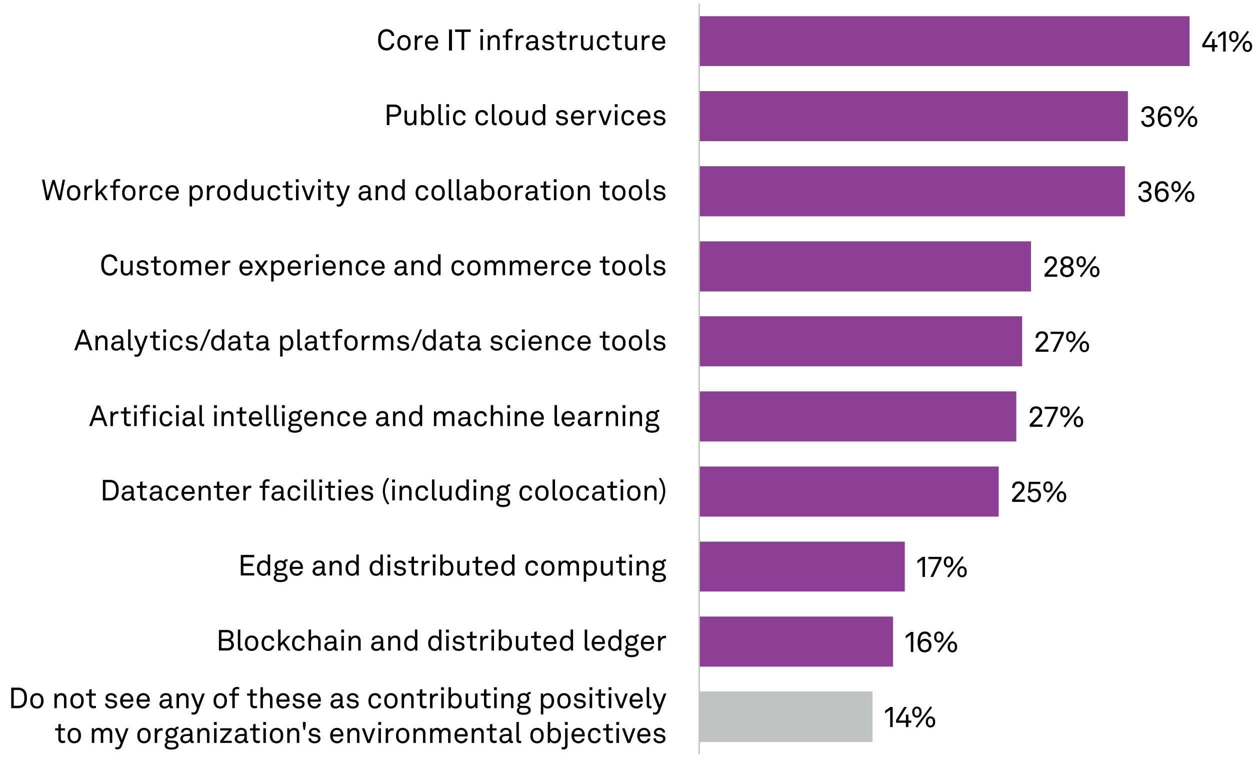
# Paths to Efficiency



Demand for digital transformation will continue to fuel significant datacenter growth for decades to come. All that relentless scaling must be offset by deploying more efficient new technologies and operating models. In our Digital Pulse survey, modernizing core IT infrastructure (cited by 41% of respondents) and adopting public cloud services (36%) are the strategies that customers are most often adopting to meet their environmental objectives. Workforce productivity and collaboration tools (36%) and customer experience and commerce tools (28%) had the next-largest impacts.

And for those further ahead in their digital transformation journeys, data platforms and data science tools (35% vs. a survey average of 27%) and AI and machine-learning tools (34% vs. 27%) were seen as the most promising contributors to their environmental efforts. In the rest of this paper, we drill down into the top two strategies — core IT infrastructure and cloud services.

### Figure 2: Top Technologies for Achieving Environmental Objectives




Q. Which of these technologies, if any, do you see as contributing positively to your organization's efforts to achieve its environmental objectives? Please select all that apply.
Base: All respondents (n=407)
Source: 451 Research's Voice of the Enterprise: Digital Pulse, Environmental Impact 2022

## Core Infrastructure

Moore's Law — the observation that CPUs effectively double in transistor count about every two years — has proven incredibly resilient since the concept was introduced nearly 60 years ago. However, simply driving up power consumption to deliver higher performance in new generations of CPUs is unsustainable. Chip companies continue to deliver energy efficiency gains as well. The familiar x86-64 architecture has been extended and iterated on for decades. And today, workload-specific accelerator chips (GPUs, TPUs, FPGAs/ASICs and quantum computers) are used as part of a heterogenous compute architecture to complement the capabilities of general-purpose "classic" CPUs.

These accelerators offer new levels of performance for the wattage consumed. In 2020 AMD said that its goal was to deliver a 30x increase in energy efficiency by 2025 for AI training and high-performance computing applications running on accelerated compute nodes. That's more than 2.5x faster than the aggregate industrywide improvement achieved over the previous five years. If realized, the 30x goal will save billions of kilowatt hours of electricity in 2025, reducing the power required for those systems to complete a single calculation by 97% over five years.

While it takes some effort to leverage these CPU and workload accelerator combinations, the cost savings offered can make it worthwhile for enterprises to adopt them from a purely budgetary standpoint. Performance-per-watt-consumed metrics should be guiding workload placement and architectural decisions for net new workloads, at least.

Power usage effectiveness (PUE) is the metric most enterprises focus on for their energy transformations. The PUE rating measures overall datacenter efficiency, including cooling, server room design, renewable energy sources, and even lighting and security. It's relatively easy to envision, and there are actions that can be taken to make immediate improvements. But more important are the efficiency levels of the servers themselves, and how well they are utilized. In the past, it has been difficult for enterprises to virtualize their servers in a way that ensures they are deployed at the highest levels of efficiency. That is changing with the availability of climate governance tools that integrate environmental data into their platforms to guide operations teams in deploying workloads based on ecological performance.

Obviously there is more that goes into workload placement decisions than just the carbon impact and energy profile. Those in highly regulated industries may not have the option to move workloads to the cloud, while others will prefer to maintain full control and responsibilities — for the performance of their infrastructure, or for the security of their data. But over time, energy efficiency concerns will work their way higher up in the decision matrix. For workloads that can't be migrated to the cloud, or otherwise need to stay on-premises, organizations should consider what upgrades could be made to gain those efficiencies in power consumption, which would ultimately serve to lower a given workload's energy use and thereby its related carbon emissions.

## Cloud Services

Centralizing IT resources using the cloud paradigm can help reduce carbon footprint compared with distributed technology. As an analogy, a bus might produce more carbon than a car, but a bus carrying 20 people will likely be more net carbon friendly than 20 cars on the road. Similarly, a pool of servers shared by thousands of applications will likely be far more cost- and carbon-efficient than thousands of servers in separate datacenters. Cloud providers enjoy huge economies of scale. We are all relying more on technology, but using cloud resources means more of us are choosing to get on the bus instead of driving alone.

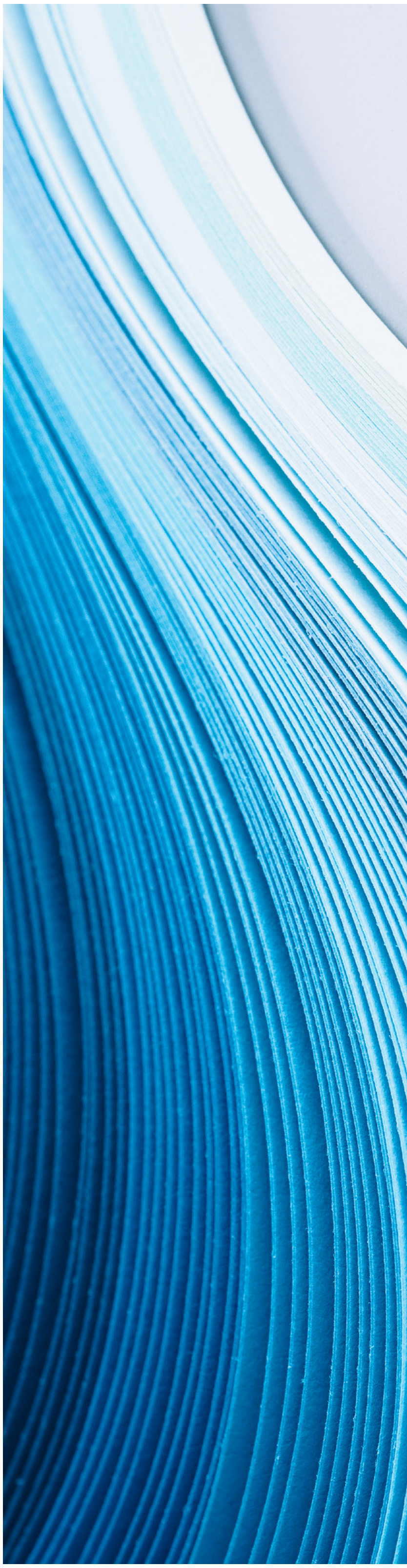
Scalability (up or down) is a major benefit of moving to cloud. Compute and storage resources can be adjusted based on usage. But for this to work effectively, applications must be broken up into virtual machines or containers that can be placed side-by-side with others to increase server utilization. Smaller units utilize spare capacity on a server more easily. This is a key tenet of cloud-native architectures, enabling applications to be managed efficiently at scale while optimizing resource usage and reducing waste.

# Conclusion

Overall, 451 Research surveys and data modeling have shown that if enterprises move their IT to the cloud, they are likely to reduce their energy usage, resulting in a smaller carbon footprint. However, those organizations still using on-premises datacenters must plan any transition carefully, making sure that operations management spans the hybrid IT estate — a mix of both on-premises and cloud-based resources — so that cloud running costs are kept to a minimum.

For efficiency, there are essentially three levers that determine the scope of savings:

- The equipment in a cloud datacenter is typically refreshed very regularly with the most cutting-edge equipment, maximizing efficiency in terms of both performance per watt and performance delivered to customers.
- Cloud infrastructure is pooled and shared for maximum utilization — and utilization is one of the biggest determinants of overall operating cost.
- The newest cloud datacenters are purpose-built for efficiency in three areas: efficient infrastructure (power and cooling); efficient servers, storage and networking; and use of clean electricity sources, renewable energy technology and renewable development projects.



Many organizations have good reasons for maintaining their on-premises datacenter operations instead of (or alongside) essential cloud services. Custom and legacy application requirements, regulatory restrictions, data sovereignty, security concerns, proximity and multicity requirements may all have to be satisfied.

For organizations in this position, the alternatives are:

- A multi-tenant facility where the PUE rating is likely to be lower than at most on-premises datacenters. Multi-tenant providers are also likely to have an aggressive sustainable energy strategy in place.
- New on-premises datacenter builds to improve PUE with more efficient infrastructure and servers — or retrofitting existing datacenters to save on the overall lifecycle emissions.

As a last resort, those emissions that remain and cannot be chased out can be offset — although this should be part of a company's continual evaluation program for seeking improvements over time. Adjustments can then be made when new products become available or old ones fail.

# About the Author


## John Abbott

**Principal Research Analyst**


John Abbott covers systems, storage and software infrastructure topics for 451 Research, a part of S&P Global Market Intelligence. Over a career that spans more than 30 years, he has pioneered specialist technology coverage in such areas as Unix, supercomputing, system architecture, software development and storage.

As one of the co-founders of The 451 Group in October 1999, John ran analyst operations from the company's San Francisco office. He has been a principal author on many 451 Research Special Reports, including those on storage virtualization and blade servers – the first comprehensive surveys of either subject to be published. More recently John has focused on topics such as converged infrastructure, new systems architectures, AI and deep learning accelerators. He helped establish 4SIGHT, the 451 Research framework for the forward-looking, long-term coverage of emerging technologies.

John began covering the technology sector in 1984, building on his previous experience as a technical author and direct involvement using mainframes, early PCs and Unix workstations. As a freelance journalist, he contributed to publications including Computing, Computer Weekly, The Financial Times and The Times. In 1987, he was appointed editor of ComputerWire's weekly Unix newsletter, Unigram.X, and later became editor of the company's daily Computergram International service, first in London and subsequently in San Francisco. He established the 451 Research office in San Francisco and lived there for over a decade.

John studied music at the University of Keele and has an MA in modern English literature from the University of London.

## About 451 Research

451 Research is a leading information technology research and advisory company focusing on technology innovation and market disruption. More than 100 analysts and consultants provide essential insight to more than 1,000 client organizations globally through a combination of syndicated research and data, advisory and go-to-market services, and live events. Founded in 2000, 451 Research is a part of S&P Global Market Intelligence.

**CONTACTS**

**The Americas**
+1 877 863 1306
market.intelligence@spglobal.com

**Europe, Middle East & Africa**
+44 20 7176 1234
market.intelligence@spglobal.com

**Asia-Pacific**
+852 2533 3565
market.intelligence@spglobal.com
www.spglobal.com/marketintelligence