

HOW MODERN ENTERPRISES ARE RETHINKING POWER ALLOCATION TO OVERCOME AI-DRIVEN CAPACITY BOTTLENECKS

WHY MEASURING HOW MUCH PROVISIONED POWER IS ALLOCATED TO COMPUTE MATTERS IN THE AGE OF AI

November 2026

Jim Rapoza
Vice President & Principal Analyst

LinkedIn

ABERDEEN
RESEARCH

Overview

In this knowledge brief, we'll look at how an emerging open metric, Power Compute Effectiveness (PCE), is enabling a next-generation method for organizations to stay competitive and sustainable.

Today businesses with large data centers find themselves at a crossroads. They need to grow in order to meet new demands for innovation and power, especially when it comes to AI. But they need to implement this growth in a way that is most effective and doesn't lead to waste and increased costs.

As AI and other emerging workloads drive rapid growth in power and cooling demands, data centers, colocation providers, service providers and large enterprises face serious capacity bottlenecks. Many are still planning with legacy metrics that show how efficiently a facility uses energy, but not how much of its provisioned power is actually allocated to compute.

With rising rack densities, more provisioned power is allocated to cooling, electrical distribution losses and auxiliary loads, leaving stranded capacity that can't support IT equipment. In this knowledge brief, we'll look at Power Compute Effectiveness (PCE), an emerging open metric defined as provisioned power for compute divided by total provisioned data center power, and how it gives organizations a better view of the way their power is allocated at the design level.

We'll also explore how data center design and cooling strategies informed by PCE can help infrastructure teams shift more provisioned power toward compute, supporting growth within existing power commitments, reducing the need for new grid capacity and improving return on invested power (ROIP), a financial outcome derived from PCE, though putting that headroom to use may still require investment in cooling and power infrastructure.

The AI Capacity Crisis: Rising Power Demands and Stranded Capacity

As much as, if not more than, any other time, modern data center organizations are working constantly to keep pace with rapid digital innovations, and these businesses are under intense pressure to remain competitive while preparing for the future. This challenge is driving them to adopt new technologies while securing a strong return on their infrastructure investments. As data center organizations rush to integrate vital technologies such as artificial intelligence, the ability to modernize quickly and cost-effectively is critical.

However, when businesses modernize their data center and infrastructure, they often find that older power architectures, cooling systems, and IT systems are creating a barrier to transformation. Looking at recent Aberdeen research into infrastructure and data center modernization, we find that:

35% The number one data center challenge, chosen by 35% of enterprise and service provider respondents, is increasing costs of resources and power.

39% A top three challenge to successful AI implementations is power usage and optimization, selected by 39% of organizations building infrastructures to support AI.

Source: Aberdeen Strategy & Research, 2026

The primary driver behind these infrastructure hurdles is the rapid shift toward increased AI capabilities. Traditional data center environments were designed for lower, uniform rack densities and when deployed in legacy facilities, high-density AI hardware creates large thermal loads that strain existing infrastructure.

To cope with these demands, facility planners often over-provision auxiliary systems, leading to a hidden penalty: as rack densities rise, a disproportionate share of provisioned power is siphoned off to run chillers, pumps, fans, electrical distribution losses, and auxiliary loads. These practices can lead to stranded capacity, with power being used for infrastructure overhead instead of valuable compute.

Aberdeen research shows that leaders in data center management and optimization (identified in our research as enterprises and service providers scoring in the top 30% of data center success metrics) are working to improve their overall data center capabilities, with our data showing that these leaders are 66% more likely than competitors to be in the process of modernizing and improving data center power capabilities, and they are 3x more likely to be modernizing cooling systems compared to their peers.

PCE: Making Stranded Capacity Visible

Developed by data center cooling and infrastructure solutions provider Airsys and accepted and published by the Open Compute Project (OCP), PCE is defined as provisioned power for compute divided by total provisioned data center power.

Key Distinctions of PCE:



Design-Level Evaluation: Unlike operational metrics that change with live IT utilization and runtime conditions, PCE is calculated during the design stage from single-line diagrams (SLDs), original equipment manufacturer (OEM) submittals, and commissioning documentation.



Effectiveness, Not Efficiency: PCE measures how provisioned power is structurally allocated across the facility. It does not measure runtime energy consumption or compute output performance.



Making Stranded Capacity Visible: By isolating how much provisioned power is dedicated to cooling rejection, electrical losses, and auxiliary loads before reaching the server, PCE helps reveal where capacity is being trapped.

By targeting an optimized PCE design goal, infrastructure teams can shift more provisioned power toward compute within the same grid power commitment and physical footprint, though using that headroom may require investment in IT hardware, cooling and power distribution.

Infrastructure Strategies for Achieving Higher PCE

Building a data center with a PCE strategic approach can require professionals to rethink data center architecture, especially where thermal management and power distribution interact. Traditional cooling architectures rely on a single, uniform loop where mechanical chillers typically process the full heat load at a low coefficient of performance (COP). This approach wastes significant provisioned capacity on auxiliary cooling overhead.

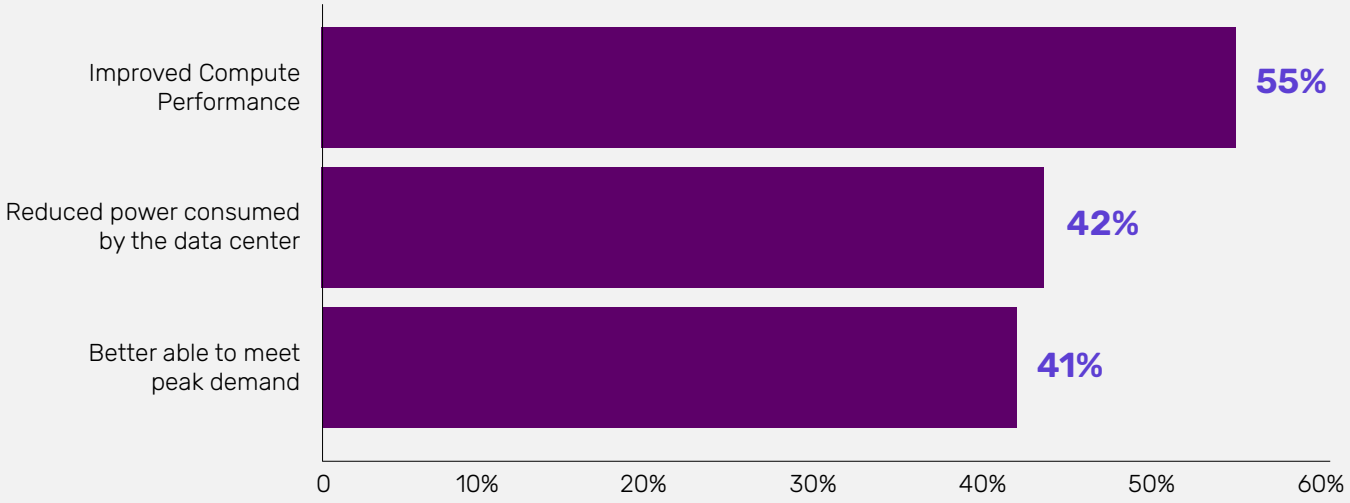
In a PCE-based infrastructure, multi-zone thermal strategies are used to match cooling loops directly to heat grades:

1. **High-Temperature (HT) Direct-to-Chip Loops:** Utilizing dry coolers and coolant distribution units (CDUs) to reject heat directly to ambient air with zero compressor energy.
2. **Medium-Temperature (MT) Air-Side Economization:** Managing residual server and ambient room heat via efficient fanwalls and air handlers.
3. **Low-Temperature (LT) Mechanical Trim:** Reserving dedicated precision cooling strictly for low-temperature electrical rooms, UPS systems, and battery energy storage system (BESS) boundaries.

By coordinating these zones with unified software logic, data centers can compress cooling rejection provisions. This shift can reallocate stranded provisioned power from cooling overhead to compute, increasing PCE; deploying IT equipment to use that power is a separate investment.

As part of an overall modernization of data center power and cooling systems, this kind of implementation can bring significant benefits to enterprises and service providers. Aberdeen research looking specifically at businesses that strategically modernized cooling systems found that these data center organizations realized important advantages, as seen in Figure 1.

Top Benefits of Modernized Data Center Cooling Strategies



Source: Aberdeen Strategy & Research, 2026

Organizations that modernized cooling reported improved compute performance, reduced data center power consumption and a better ability to meet peak demand. These are self-reported outcomes of cooling modernization broadly. PCE complements them by showing, at the design stage, how much provisioned power is allocated to compute rather than to cooling, electrical losses and auxiliary loads.

Key Takeaways

As AI workloads continue to scale, enterprises and service providers can no longer rely solely on operational energy-efficiency metrics such as PUE. True infrastructure modernization requires total visibility into how power is provisioned and allocated from the ground up.

By adopting Power Compute Effectiveness (PCE) as a core design standard, enterprises, colocation providers, and data center operators can overcome AI-driven capacity bottlenecks. Aligning architecture around PCE enables businesses to allocate more of their provisioned power to compute within existing grid commitments, reduce the need for new grid capacity and improve return on invested power (ROIP), the financial outcome that results when that reallocated power supports revenue-generating IT relative to the infrastructure investment required, paving the way for sustainable, high-density AI growth.

About Aberdeen Strategy & Research

Aberdeen Strategy & Research, a division of Spiceworks Ziff Davis, with over three decades of experience in independent, credible market research, helps illuminate market realities and inform business strategies. Our fact-based, unbiased, and outcome-centric research approach provides insights on technology, customer management, and business operations to inspire critical thinking and ignite data-driven business actions.

This document is the result of primary research performed by Aberdeen and represents the best analysis available at the time of publication. Unless otherwise noted, the entire contents of this publication are copyrighted by Aberdeen and may not be reproduced, distributed, archived, or transmitted in any form or by any means without prior written consent by Aberdeen.