InfoBites, sponsored by Supermicro and NVIDIA | JULY 2025

Powering AI Factories: Scaling GenAI with Direct-to-Chip Liquid-Cooling

Transforming Datacenters from Cost Centers to Engines of Intelligence

Andrew Buss
IDC IT infrastructure EMEA

Luis Fernandes
IDC IT infrastructure EMEA

GenAI datacenters will need new approaches to maximize efficiency.

Key infrastructure-related challenges in building out GenAI infrastructure seen by Digital Leaders:[1]

35%
Deploying GenAI to remote datacenters for power or ESG reasons

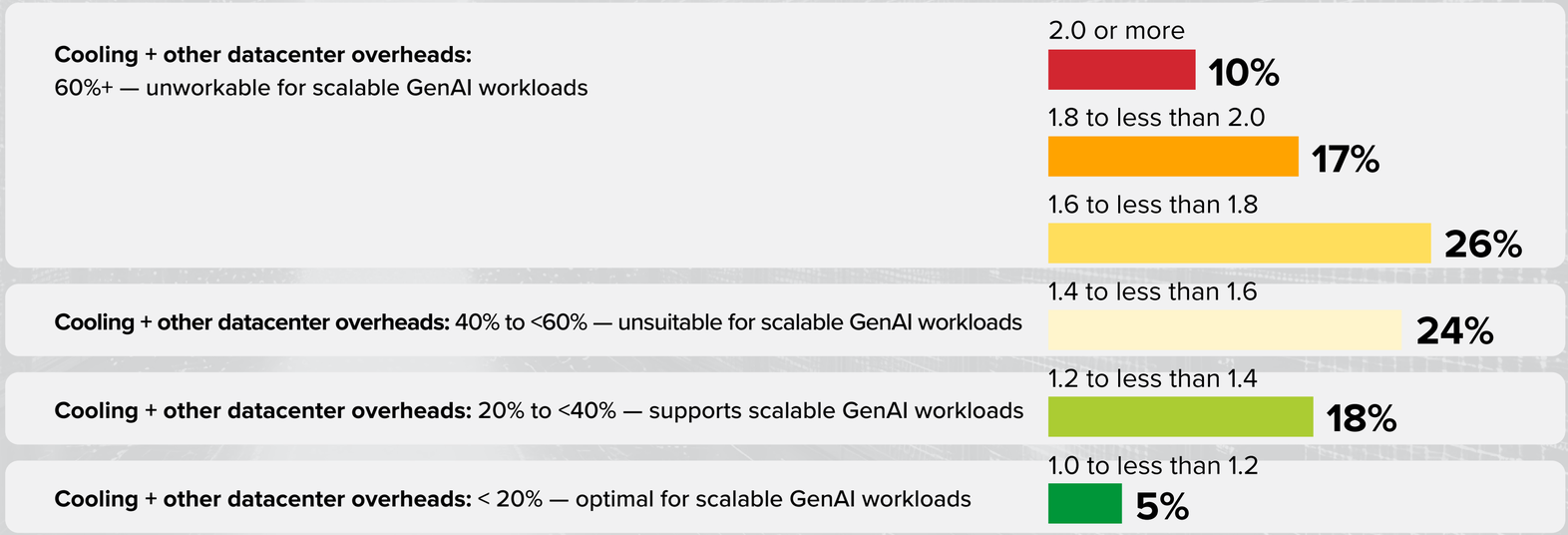
34%
Significantly increased power consumption within IT infrastructure

29%
The rising cost of power required to run GenAI workloads

17%
Not getting enough power to run GenAI workloads at the scale we need

The majority of air-cooled datacenters in operation today waste too much energy on cooling to run GenAI infrastructure effectively at scale.

The PUE of an Organization's Most Efficient Datacenter in 2023



To support **scalable and sustainable GenAI solutions**, we need to **maximize the energy directed toward powering GPUs and AI accelerators** while **minimizing the energy consumed by cooling, power distribution**, and other datacenter functions that constitute overhead rather than contributing direct value.

©IDC 2025 | IDC #EUR153593925

Sources: 1) IDC's *AI-Ready Infrastructure Survey, 2024*, EMEA (n = 1,590); 2) IDC's *Edge to Cloud Infrastructure Survey, 2023*, EMEA (n = 1,626)

InfoBites, sponsored by Supermicro | NVIDIA

Message from the sponsor

Supermicro and NVIDIA are redefining the economics of deploying AI factories. We offer state-of-the-art infrastructure solutions that address increased power and cooling challenges in modern AI datacenters. Additionally, significant savings can be achieved with direct liquid-cooling (DLC-2) for highly efficient generative AI datacenters.

For more information

©IDC 2025 | IDC #EUR153593925