InfoBites, sponsored by Supermicro and NVIDIA | JULY 2025

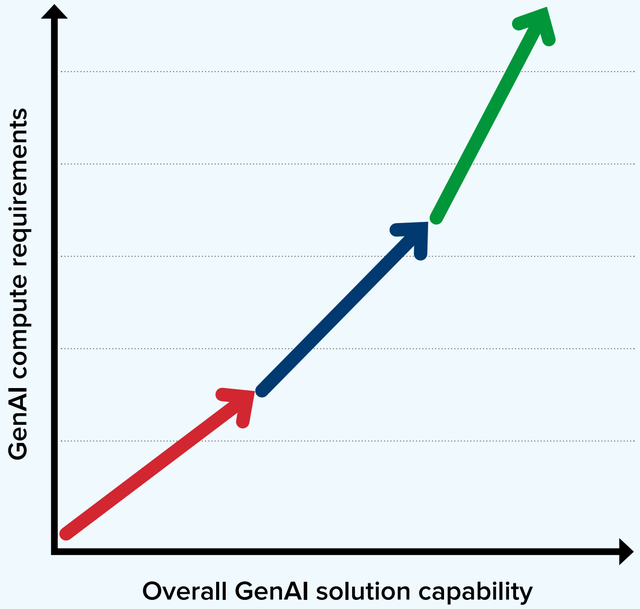
Powering AI Factories: Scaling GenAI with Direct-to-Chip Liquid-Cooling

Transforming Datacenters from Cost Centers to Engines of Intelligence

Andrew Buss
IDC IT infrastructure EMEA

Luis Fernandes
IDC IT infrastructure EMEA

Advances in GenAI reasoning will dramatically increase the need for accelerated computing.



Test-time scaling: Instead of processing a one-time answer, models allocate extra computational effort during inference, reasoning through multiple responses before arriving at an optimized answer.

Post-training scaling: The performance of a pretrained model can be improved by using techniques such as fine-tuning, distillation, pruning, quantization, reinforcement learning, and synthetic data augmentation.

Pre-training scaling: Increasing training dataset size, model parameter count, and computational resources results in predictable improvements in model intelligence and accuracy.

AI scaling laws are **driving exponential compute demand**. Establishing AI factories is essential in enabling this emerging model, in much the same way that foundational infrastructure was once required for the widespread adoption of electricity and the internet.

To support AI reasoning and agentic AI, test-time scaling can require up to 100 times more compute than standard inference. This is already having a dramatic impact with significant increases for power and cooling requirements for AI datacenters and infrastructure.

Source: IDC's *AI-Ready Infrastructure Survey, 2024*, EMEA (n = 1,590)

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