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

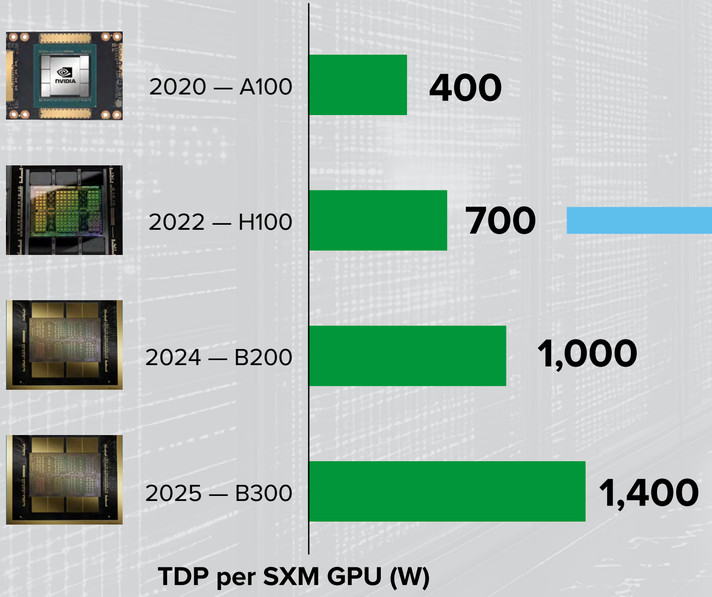
Andrew Buss
IDC IT infrastructure EMEA

Luis Fernandes
IDC IT infrastructure EMEA

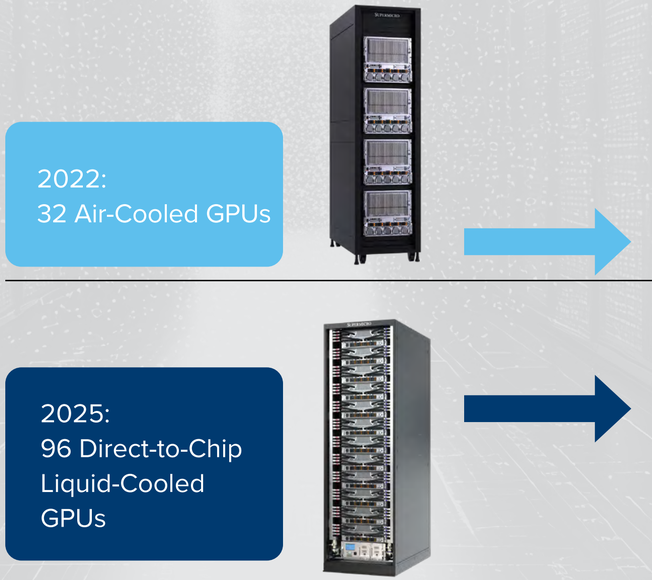
Direct-to-chip liquid-cooling enables the leap from traditional datacenters to high-throughput AI factories.

The performance needs of GenAI mean that rack scale systems are being engineered for maximum compute performance and density. Even with significant increases in performance per watt, increasing demand for scalable GenAI compute still means that overall GenAI compute power density is rising rapidly.

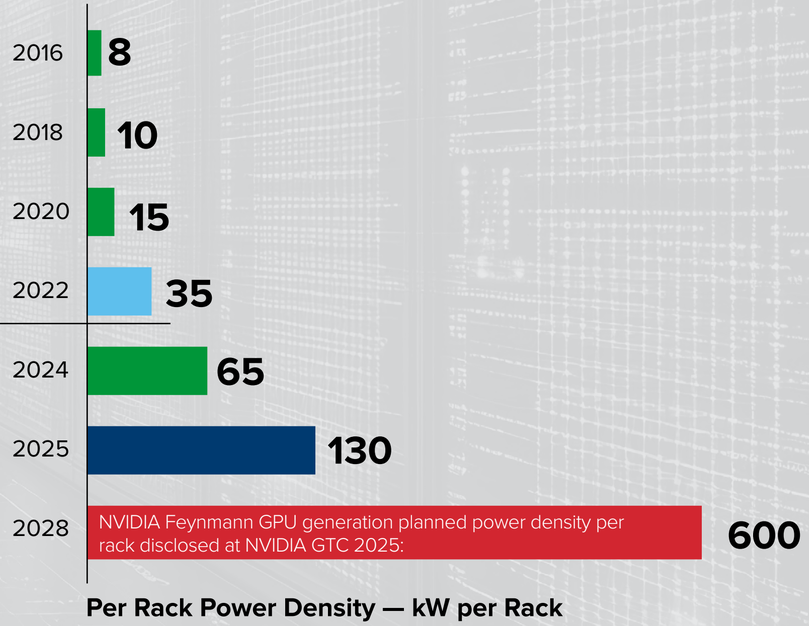
NVIDIA Datacenter GPUs TDP per SXM GPU (W)



Datacenter GPUs per Rack



kW per Rack for GPU-Accelerated GenAI Compute



Datacenters are limited by the available power from the utility provider. That means that revenues are becoming power limited. Power efficiency is a central metric for operational and revenue success; every watt not used for AI inference or training is lost revenue.

Sources: NVIDIA product specifications; NVIDA "GTC" keynote, March 2025; Wade Vincent's "Data Center World" keynote, April 2025; IDC

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