



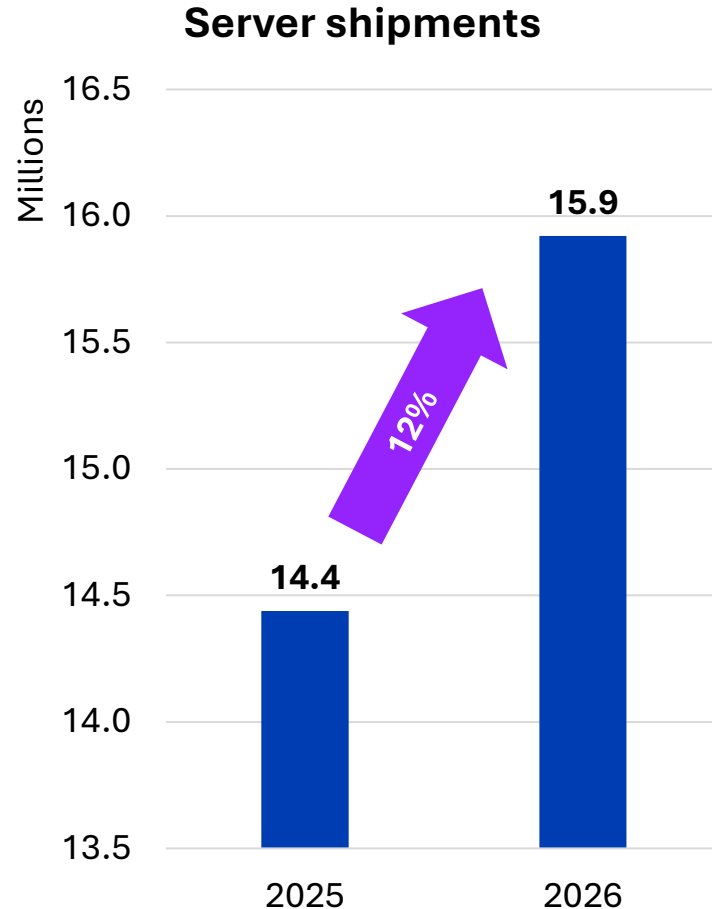
Cloud and Data Center Market Snapshot – February 2026

Vladimir Galabov

Senior Research Director, Enterprise Infrastructure

askananalyst@omdia.com

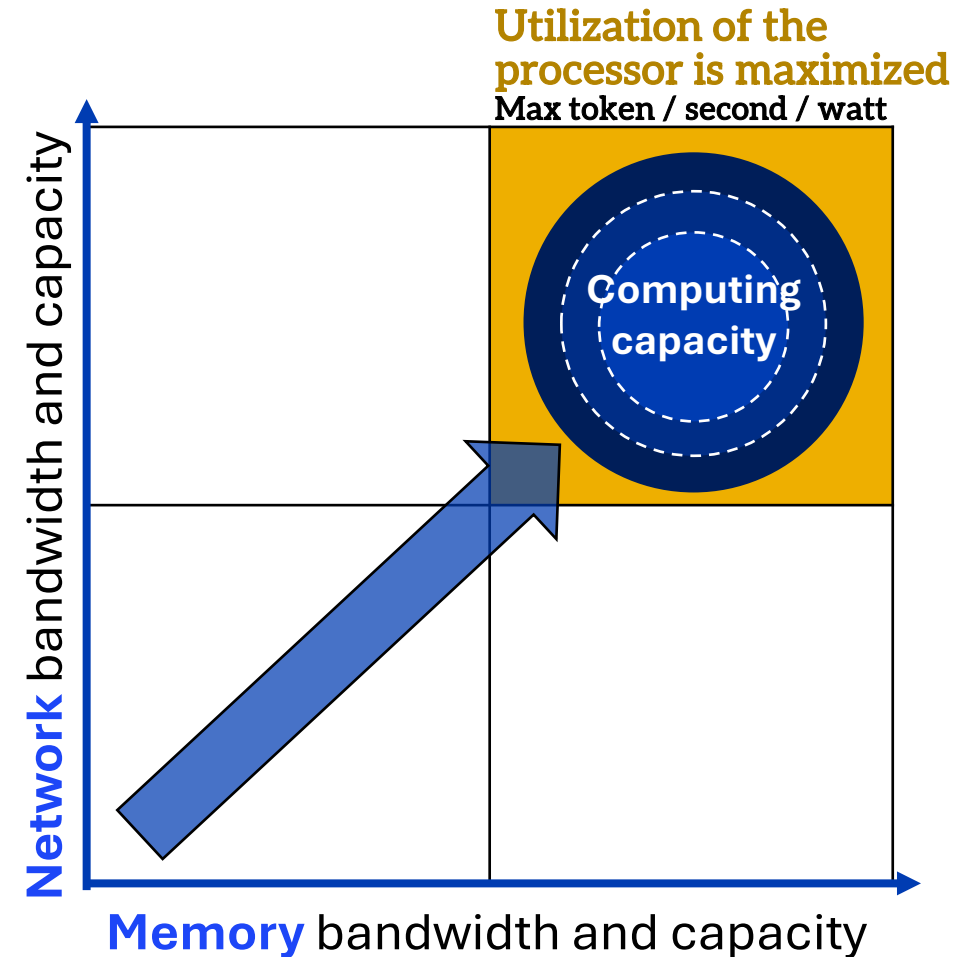
An IT supply chain in flux



- **We are entering 2026 in an increasingly challenging IT supply chain situation.**
- **We're in a memory super cycle. Memory availability is very poor and prices are up 70% year-over-year.**
 - DRAM is in particularly short because of the extraordinary AI demand and because memory vendors repositioned capacity towards HBM.
 - Beyond escalating costs, we are concerned about memory availability disrupting server manufacturing which in turn disrupts the ability for the data center industry to complete ongoing projects. For now, servers are prioritized ahead of PCs and smartphones and should be less impacted by memory shortages.
- **We can confirm that data center CPUs are also experiencing supply constraints.**
 - As CPU roadmaps have broadened vendors are simultaneously manufacturing CPUs on different process nodes - e.g., 5nm and 3nm. Vendors claim that they are struggling to match demand with manufacturing capacity across different process nodes. The nature of processor design and fabrication is such that shifting volume across process nodes is complex and time consuming. Lower than expected yields could have also contributed to this shortage.
 - Likely manufacturing line allocation has favored the high-value AI chips ahead of CPUs at TSMC.
 - We expect that CPU prices could increase in 11-15% in 2026 because of the shortage. This conservative estimate is driven by the fact that hyperscale cloud service providers sign long term agreements with their vendors resulting in somewhat fixed prices for the upcoming year - a standard business practice.
- **We continue to forecast 12% growth rate in server shipments as the refresh cycle we highlighted last year continues. We do not see a risk to our forecast from CPU availability, but we do see a real risk from memory availability.**

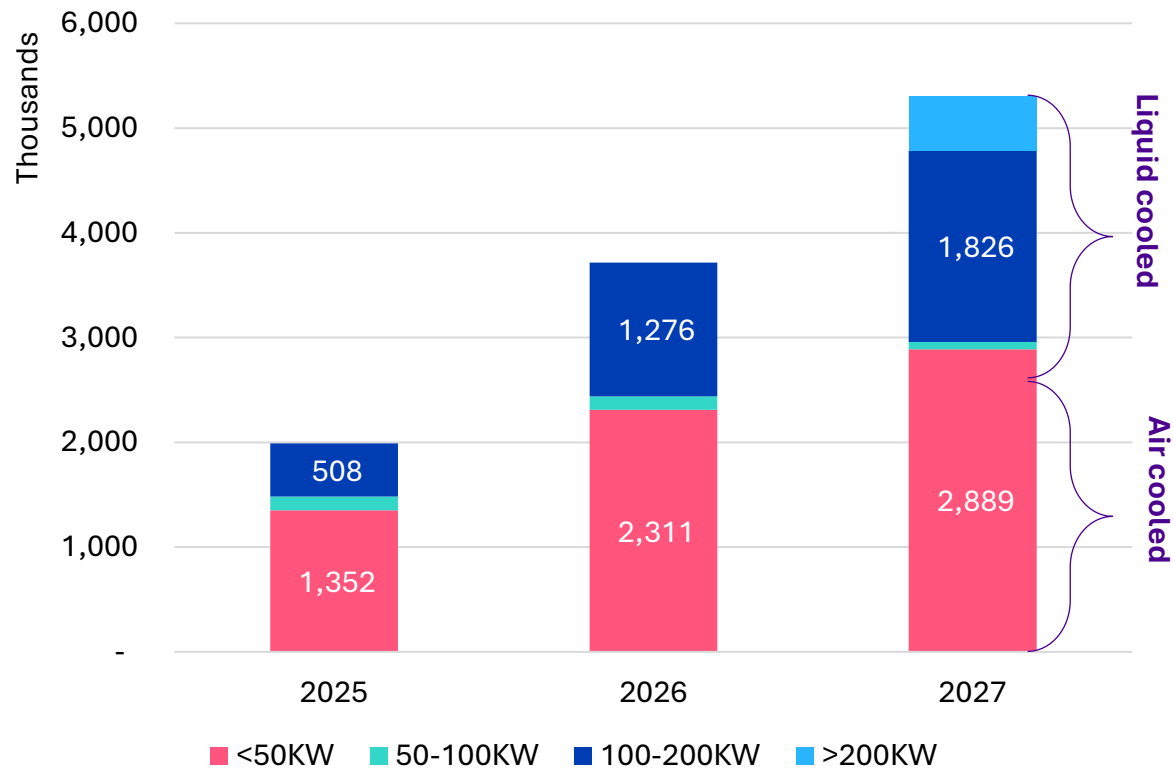
Why rack power density continues to increase?

- To maximize the token per dollar per watt system performance system performance has to be optimized again three parameters:
 1. Network bandwidth and capacity
 2. Memory bandwidth and capacity
 3. Computing capacity of the GPU or ASIC
- Copper based scale-up network interconnects continue to provide the best throughput and power efficiency. However, their performance is restricted by distance. Copper cables can reach less than 2m distance.
 - The fast and efficient network enables maximum utilization of each GPU, i.e. computers do not sit idle because of a network bottleneck.
 - 70% or more of the power consumption of a rack-scale system is for data transfer, not processing.
- This consideration is what is driving the growth in rack density across NVIDIA, AMD and Huawei rack-scale AI-optimized designs.
- We see AWS, Microsoft and Google developing less dense designs and, in the process, gaining operational and maintenance benefits.

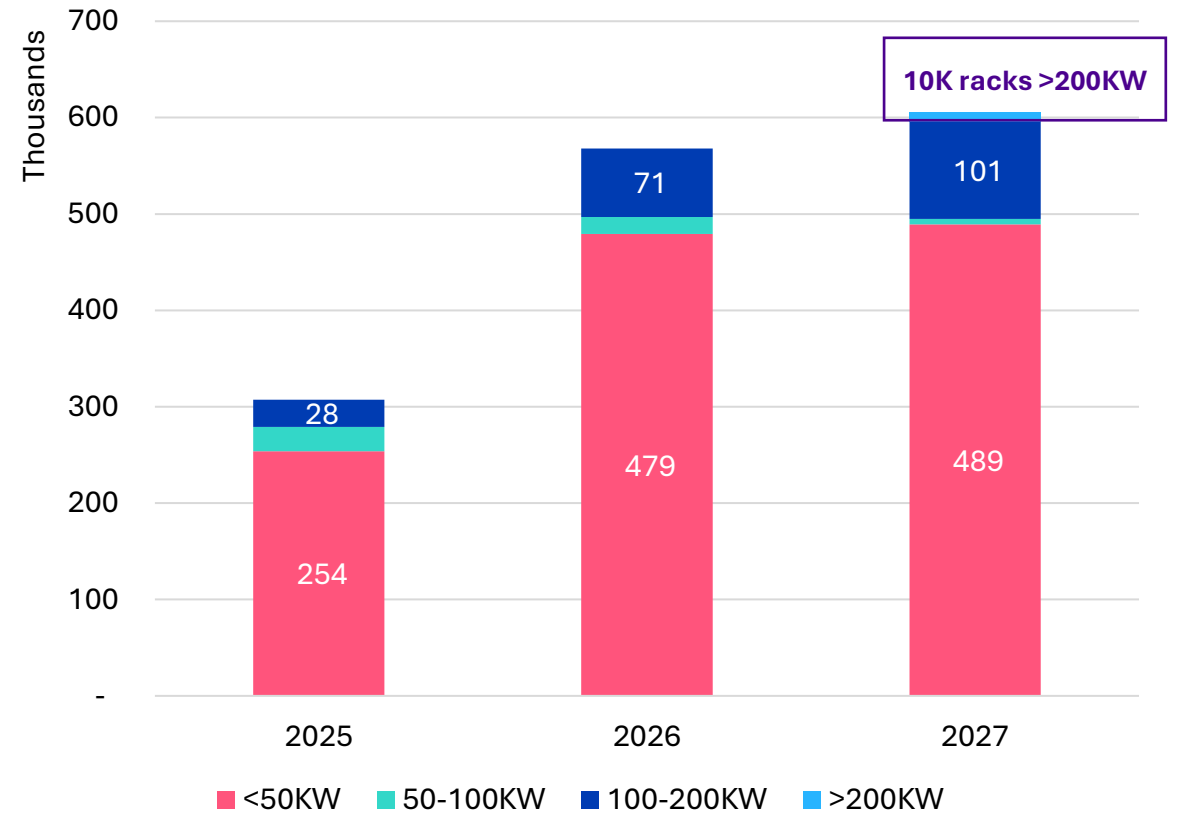


71K racks with > 100KW capacity will ship out this year, the vast majority NVIDIA NVL72
 This will increase by 56% next year and we will see the first ultra dense racks (> 200KW)

AI-optimized servers split by the power capacity of the rack they're installed in



AI-optimized racks split by their power capacity



Join the debate!

- The team which brings you this market snapshot has a new free weekly podcast.

Episode 1: Trends and Truths

- Part One: Truths – level setting and myth busting
- Part Two: Trends – 2026 trends to watch
 - 00:00 Memory and storage super cycle - observations and drivers
 - 05:45 Ramp of rack scale AI systems - shipments, vendors and users
 - 09:25 Networking innovation and co-packaged optics deployments
 - 10:55 Data center energy autonomy and self-generation projects
 - 14:55 High voltage power distribution in the data center
 - 17:45 Battery energy storage systems deployments in the data center
 - 19:09 Two-phase direct-to-chip technology benefits and market formation
 - 21:45 Single-phase direct-to-chip cold plate innovation
 - 22:10 Most likely approaches to microfluidic liquid cooling

Episode 2: The Death of the Metaverse



Disclaimer

The Omdia research, data and information referenced herein (the “Omdia Materials”) are the copyrighted property of TechTarget, Inc. and its subsidiaries or affiliates (together “Informa TechTarget”) or its third party data providers and represent data, research, opinions, or viewpoints published by Informa TechTarget, and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa TechTarget does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an “as-is” and “as-available” basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa TechTarget and its affiliates, officers, directors, employees, agents, and third party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa TechTarget will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.

Get in touch

Americas

customersuccess@omdia.com

08:00 – 18:00 GMT -5

Europe, Middle East & Africa

customersuccess@omdia.com

8:00 – 18:00 GMT

Asia Pacific

customersuccess@omdia.com

08:00 – 18:00 GMT + 8