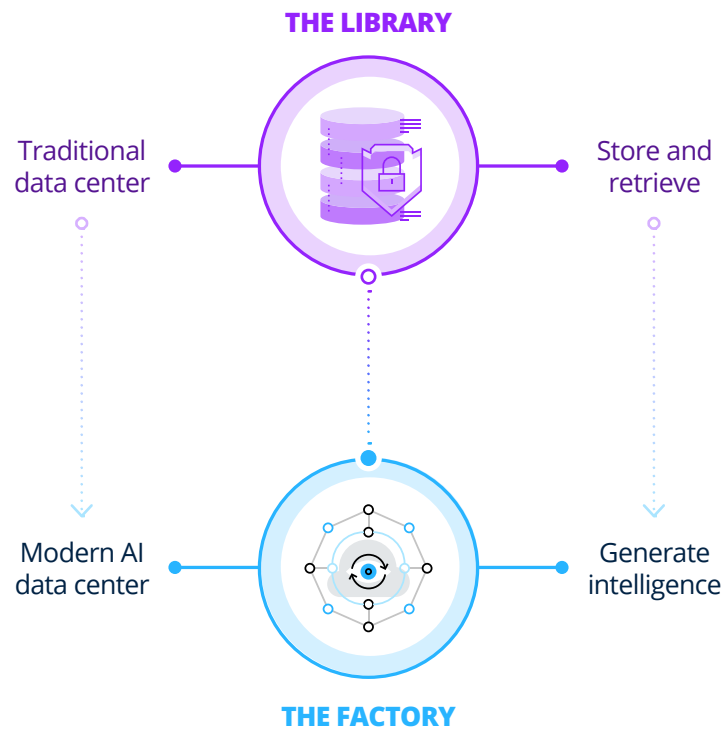
The AI Factory Infrastructure Transformation

How enterprises can move from buying servers to generating intelligence – efficiently, at scale, and with certainty of outcome.

The Compute Revolution

The rise of agentic AI in the enterprise demands a change in the compute platform - from a library of data to a factory of intelligence. As AI moves from pilot to production, the infrastructure underneath must evolve from a cost center into an intelligence engine.



The $870B Reality

Omdia analysis shows Enterprise AI spend is scaling faster than any other previous technology, and accelerated computing will therefore represent ~83% of the $870B data center server market by 2030.

83%

The Deployment Challenge

Many AI initiatives struggle to scale when compute, networking, storage, cooling, software, site readiness, and operating ownership are planned independently. High utilization and faster value realization depend on integrating and validating the complete system.

The rewards of success: Of enterprises who have scaled AI:

22%

say they've seen a **10-20% increase in revenue**

11%

say top-line revenue has increased by **more than 20%**

Turnkey Solutions Solve Execution Problems

Turnkey AI factory solutions combine validated reference architectures with global manufacturing capabilities and complete system integration. These solutions deliver production-ready systems from 8 to 1,000+ GPUs, eliminating the execution risks that cause most AI projects to fail.

Time-to-online:

From signed PO to production tokens in weeks rather than quarters - measured against the typical DIY timeline of 6–12 months and frequent slip-to-production scenarios.

Software stack:

GPU orchestration and scheduling, integrated software for cluster operations, and digital twin software platform.

Operational tooling:

Cluster management, monitoring, observability tooling, and support contracts from both partners - covering hardware, software, and stack interactions end-to-end.

Cooling:

Rack-level Direct Liquid Cooling (DLC) for highest efficiency and compute density, or air cooled optimized racks for constrained sites without liquid cooling.

Networking:

Pre-cabled, pre-validated, and benchmarked against the reference architecture's non-blocking bandwidth specification.

Integration & validation:

L11 (rack-level) and L12 (multi-rack cluster-level) validation reports, on-site installation, and global service from suppliers.

Compute:

Selected to match workload profile and density requirements.

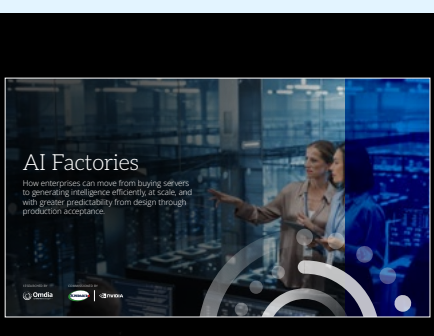
Storage:

Optional integrated storage systems supporting the unique requirements of the workload - sized to keep GPUs fed at production load rather than starved on I/O.

Learn more

Download the eBook here

Supermicro Data Center Building Block Solutions® combine NVIDIA accelerated computing with workload-matched compute, networking, storage, cooling, management software, rack integration, validation, and deployment services.



Omdia
Omdia.com
Omdia
OmdiaHQ

Supermicro
Supermicro.com
Supermicro
Supermicro

NVIDIA
NVIDIA.com
NVIDIA
NVIDIA

COPYRIGHT © 2026 OMDIA TECHTARGET