LIQUID COOLING — GUIDE

Rack Density Trends for AI: From 30 kW to 150+ kW per Rack

Track how AI rack power densities have escalated from 30 kW to 150+ kW, driven by NVIDIA GB200 NVL72, AMD Helios, and hyperscale deployment data.

Updated November 1, 2026

Rack power density is the defining constraint for AI data center design. This guide tracks the escalation from 30 kW to 150+ kW per rack and what it means for cooling infrastructure.

Density timeline

Year	Typical peak AI rack	Driver	Cooling implication
2020-2021	15-25 kW	V100, A100 (400W)	Air cooling with containment
2022	20-30 kW	A100 80GB, H100 (700W)	Air cooling at limit, RDHx adoption
2023	30-50 kW	H100 8-GPU HGX, H200	Direct-to-chip pilots, CDU deployment
2024	70-100 kW	B200 (1,000W), GB200 NVL36	Liquid cooling standard for new AI
2025-2026	100-140+ kW	GB200 NVL72, Helios, Vera Rubin	Full liquid-cooled rack architectures

Sources: TrendForce (2024); NVIDIA GB200 NVL72 specs (2026); Mordor Intelligence (2026); Uptime Institute 2026 Global Survey (cited in Mordor Intelligence, 2026).

Key density milestones

NVIDIA GB200 NVL72: ~140 kW per rack

- 36 Grace CPUs + 72 Blackwell GPUs in one rack
- 72-GPU NVLink domain, 130 TB/s interconnect
- Direct-to-chip liquid cooling mandatory
- Facility water loop via CDU; closed-loop design (NVIDIA, 2026; TrendForce, 2024)

AMD Helios: Rack-scale AI reference

- Quick-disconnect liquid cooling integrated in rack design
- Targeted at gigawatt-scale deployments (July 2026 production)
- OEM/ODM customizable reference (Mordor Intelligence, 2026)

NVIDIA Vera Rubin DSX: Partner guidance for power/cooling/control

- Full infrastructure stack reference including cooling
- Announced March 2026 (Mordor Intelligence, 2026)

Google TPU Pods: 7+ years at gigawatt scale

- Liquid-cooled TPU deployments across 2,000+ pods since ~2018
- Demonstrates operational maturity at scale (Mordor Intelligence, 2026)

Density distribution in practice

Uptime Institute's 2026 global survey (cited in Mordor Intelligence, 2026):

- **Average modal density**: Slow upward trend (~8-12 kW/rack industry-wide)
- **Peak density for AI/HPC**: More operators reporting "e30 kW/rack for advanced workloads
- **Key insight**: Facility-wide average is a poor guide; peak rack heat loads justify targeted liquid cooling investment

Cooling technology by density band

| Density band | Viable cooling approaches | Notes | |--------------|--------------------------|-------| | 15-30 kW | Air (contained), RDHx | Air still viable with good containment | | 30-50 kW | RDHx, direct-to-chip | Direct-to-chip preferred for new builds | | 50-100 kW | Direct-to-chip, immersion | Air not viable; CDU required | | 100-150 kW | Direct-to-chip, two-phase immersion | RDHx insufficient; full liquid loop | | 150+ kW | Two-phase immersion, advanced direct-to-chip | Reference architectures only (2025-2026) |

Implications for facility design

Power infrastructure

- 140 kW rack = ~195 A at 415V 3-phase or ~168 A at 480V 3-phase
- Busway and breaker sizing must accommodate peak + diversity
- DC distribution (380V DC, ±400V DC) under evaluation for >100 kW racks

Cooling infrastructure

- **CDU capacity**: Row-based CDUs (300 kW - 2 MW) for multi-rack pods
- **Facility loop**: Dry coolers for W27+ classes; chillers only for W17
- **Piping**: Larger diameters, insulated supply/return, quick-disconnects at rack
- **Redundancy**: N+1 CDU minimum; 2N for mission-critical

Structural

- 140 kW rack + CDU + piping = 1,500-2,500 kg per rack position
- Load spreaders often required for brownfield retrofits (Mordor Intelligence, 2026)
- Seismic bracing for liquid-filled piping

Deployment models

- **New-build AI factory**: Designed for 100-150 kW/rack from day one. Google, Microsoft, Oracle, AirTrunk, Equinix building to this spec (Mordor Intelligence, 2026).

- **Brownfield retrofit**: Incremental. Replace air-cooled row with single high-density liquid rack + liquid-to-air CDU. IEA 4E (2026) notes complete overhauls often impractical.
- **Colocation**: Hall-ready for liquid cooling (Teraco JB4, Equinix KL2) enables tenant deployment without facility overhaul.

Planning recommendations

1. **Design for the peak rack, not the average**. A 50 kW average with one 140 kW rack needs liquid cooling for that rack.
2. **Specify W-class early**. W27-W32 enables free cooling; match to OEM server spec.
3. **Size CDU for roadmap**. A 100 kW rack today may become 150 kW in 18 months.
4. **Qualify fluids with OEM**. Intel certified Shell immersion fluids for Xeon (Mordor Intelligence, 2026); NVIDIA/AMD reference designs specify coolant requirements.
5. **Plan for refrigerant phase-down** (if using two-phase). EU F-gas, US AIM Act affect fluorocarbon supply.

Sources

- TrendForce. "NVIDIA Blackwell's High Power Consumption Drives Cooling Demands." 30 July 2024. https://www.trendforce.com/presscenter/news/20240730-12232.html
- NVIDIA. "NVIDIA GB200 NVL72 Specifications." 2026. https://www.nvidia.com/en-us/data-center/gb200-nvl72/
- Mordor Intelligence. "Data Center Liquid Cooling Market Report." 2026. https://www.mordorintelligence.com/industry-reports/data-center-liquid-cooling-market
- Wikipedia. "Power usage effectiveness." Last edited 17 October 2026. https://en.wikipedia.org/wiki/Power_usage_effectiveness
- IEA 4E. "EDNA Report on Data Center Cooling." 2026. (Cited in Mordor Intelligence, 2026)

Frequently asked questions

What is the current peak rack density for AI clusters?

NVIDIA GB200 NVL72 racks reach ~140 kW. AMD Helios and NVIDIA Vera Rubin DSX reference designs target similar densities. Uptime Institute's 2026 survey found more operators reporting peak densities of 30 kW or higher for advanced workloads (Mordor Intelligence, 2026).

Is 150 kW per rack typical or exceptional?

150 kW is the leading edge for 2025-2026 deployments. Most AI clusters today run 50-100 kW/rack. The 140-150 kW designs (GB200 NVL72, Helios) are reference architectures for hyperscale new-builds.

How fast are rack densities growing?

Average modal densities rise slowly, but peak densities for AI workloads jumped from ~20 kW (2022) to ~50 kW (2023) to 100-140 kW (2024-2025) with each GPU generation (H100 !' H200 !' B200/GB200).

Do I need liquid cooling at 30 kW?

Air cooling can handle 30-40 kW with advanced containment and high airflow, but liquid cooling becomes more reliable and efficient above 30 kW. Uptime Institute links rising peak densities to AI applications (Mordor Intelligence, 2026).