AI Data Center Glossary

Plain-language explanations of the terms you'll run into when reading about AI data centers — the massive facilities being built to train and run today's AI systems.

Compiled by Power Grab TX — powergrabtx.com

AI & Compute Basics

The concepts behind what these data centers are actually built to do.

Artificial Intelligence (AI)

Software that performs tasks — recognizing images, writing text, making predictions — that normally require human judgment. Modern AI mostly refers to systems trained on huge amounts of data rather than programmed with explicit rules.

Machine Learning

The technique behind most modern AI: instead of writing step-by-step instructions, engineers show a computer program millions of examples and let it work out the patterns for itself.

Large Language Model (LLM)

An AI system trained on huge amounts of text so it can generate and understand language — the technology behind chatbots like Claude, ChatGPT, and Gemini.

Neural Network

The general structure most AI models are built from — layers of simple mathematical units loosely inspired by neurons in the brain, connected together to find patterns in data.

Parameters

The internal “knobs” a model adjusts while learning — often numbering in the billions. Roughly speaking, more parameters mean a more capable but more computationally expensive model.

Think of them as the settings a giant mixing board would need to reproduce a specific sound — an AI model has billions of tiny settings tuned during training.

Training

The process of “teaching” a model by having it process massive datasets and adjust its parameters over and over. This is the most computationally intensive part of AI and the main reason new, huge data centers are being built.

Inference

Using an already-trained model to actually answer a question or generate a response — what happens every time you send a message to a chatbot. Training happens once (or periodically); inference happens constantly, at massive scale, for every user.

Compute

Shorthand for raw processing power — the amount of computing capacity available to do a task. When people say “AI needs more compute,” they mean more chips, more electricity, and more data centers.

Cluster
A large group of computers (or chips) wired together and coordinated to work on one job as if they were a single, much more powerful machine.

Supercomputer
An extremely powerful computing system built from thousands of interconnected chips. Many of today's “AI supercomputers” are really just very large, purpose-built GPU clusters.

Chips & Hardware
The physical components that do the actual computing.

GPU (Graphics Processing Unit)
A chip originally designed to render video game graphics that turns out to be extremely good at the type of math AI training needs. GPUs (mainly from Nvidia and AMD) are the single biggest driver of AI data center demand and cost.

CPU (Central Processing Unit)
The traditional “brain” of a computer, good at doing tasks one after another very quickly. CPUs still run the general software in a data center, but they're too slow for large-scale AI training compared to GPUs.

TPU (Tensor Processing Unit)
Google's custom-designed chip, built specifically for AI workloads instead of general computing or graphics. It's a competitor/alternative to Nvidia's GPUs.

Accelerator
An umbrella term for any chip — GPU, TPU, or other custom design — built to speed up AI workloads specifically, as opposed to general-purpose computing.

ASIC (Application-Specific Integrated Circuit)
A chip custom-designed for one narrow job rather than general use. Some companies build their own AI ASICs to reduce reliance on Nvidia and cut costs at scale.

Semiconductor / Chip
The physical piece of silicon that hardware like GPUs and CPUs is built on. “The chip shortage” and “chip war” headlines are about the global supply of these components.

Server
A computer designed to be housed in a data center and run continuously, providing computing power or storage over a network rather than being used directly by one person.

Rack
A standardized metal frame that servers slide into, stacked vertically, so a data center can pack many machines into a small footprint. AI racks are far denser and hungrier for power than traditional ones.

Node
One individual computing unit within a larger cluster — often a single server containing several GPUs.

HBM (High Bandwidth Memory)

A specialized, very fast type of memory stacked directly onto or next to a GPU so it can feed the chip data quickly enough to keep it busy. It's also one of the most supply-constrained components in the current AI chip market.

Power & Energy

AI data centers are, first and foremost, enormous consumers of electricity — much of the current conversation about them is really about power.

Megawatt (MW) / Gigawatt (GW)

Units used to describe how much electricity a data center draws. One megawatt can power roughly a thousand average homes; large AI data center campuses are now being described in gigawatts (thousands of megawatts) — comparable to the output of a full-sized power plant.

Power Purchase Agreement (PPA)

A long-term contract where a data center operator agrees to buy a fixed amount of electricity from a power producer (a wind farm, solar farm, or even a nuclear plant) for years at a set price — used to lock in supply for facilities that need power around the clock.

Grid Interconnection

The process of physically and administratively connecting a new data center to the regional electricity grid. Long waiting lists for interconnection are one of the biggest bottlenecks slowing down new AI data center construction today.

Substation

A facility that steps electricity down from high-voltage transmission lines to a voltage a data center can actually use — essentially the plug between the power grid and the building.

Behind-the-Meter Power

Electricity generated on-site (say, from a dedicated gas plant or small reactor next to the data center) and used directly, without passing through the public grid — a workaround some operators are using to get power faster than grid connections allow.

UPS (Uninterruptible Power Supply)

A large battery system that instantly kicks in if grid power cuts out, keeping servers running for the seconds or minutes it takes backup generators to start.

Backup Generator

Typically diesel or gas engines on-site that can power the entire facility if the electrical grid fails, kept ready to run for hours or days.

Redundancy (e.g. “N+1”)

Building in spare capacity — extra generators, cooling units, or power feeds beyond what's strictly needed — so that if one piece of equipment fails, there's a backup ready and nothing goes down.

PUE (Power Usage Effectiveness)

A measure of data center efficiency: total facility power divided by the power actually used by the computers themselves. A PUE of 1.0 would mean zero energy wasted on cooling, lighting, and other overhead; well-run facilities today aim for around 1.1–1.3.

Cooling

AI chips run extremely hot and densely packed, which has made cooling technology a major topic on its own.

Air Cooling

The traditional method: large fans and air conditioning units push cool air across servers to carry heat away. It's cheap and simple but increasingly can't keep up with how much heat dense AI hardware produces.

Liquid Cooling

Using fluid instead of air to remove heat, since liquid carries heat away far more efficiently. This has become close to essential for high-end AI hardware.

Direct-to-Chip Cooling

A liquid cooling method where coolant runs through small pipes or plates that sit directly on top of the hottest chips, pulling heat away right at the source rather than cooling the whole room.

Immersion Cooling

An even more intensive approach where entire servers are submerged in a special, non-conductive fluid that absorbs heat directly from every component at once.

WUE (Water Usage Effectiveness)

A measure of how much water a data center consumes (often for evaporative cooling) relative to the computing power it delivers — a growing concern in drought-prone regions where new data centers are being built.

Waste Heat Reuse

Capturing the heat a data center generates and putting it to use elsewhere — for example, piping it to warm nearby homes or greenhouses — instead of simply venting it outside.

Networking & Connectivity

How thousands of chips are wired together so they can act as one machine, and how data moves in and out.

Interconnect

The wiring and switching technology that links chips, servers, and racks together. In AI clusters, how fast and cleanly chips can talk to each other often matters as much as how fast any single chip is.

Bandwidth

How much data can move through a connection at once — like the width of a pipe. Higher bandwidth means more information can flow between chips or across the internet per second.

Latency
The delay between sending data and it arriving. Low latency matters enormously for AI training, where thousands of chips must constantly exchange small amounts of data in near-perfect sync.

InfiniBand / NVLink
Two specific high-speed connection technologies (InfiniBand is an industry standard; NVLink is Nvidia's own) used to link GPUs together far faster than typical internet or office networking hardware could manage.

Network Switch
A device that directs data traffic between servers, racks, and rows within a data center — the traffic-control layer that keeps thousands of machines from talking over one another.

Fiber Optic Cable
Cabling that sends data as pulses of light rather than electricity, allowing much faster and longer-distance data transmission — used both within data centers and to connect them to the wider internet.

Facilities & Business Terms
Who builds and owns these facilities, and how the industry talks about them.

Hyperscaler
One of the handful of companies operating at massive, global scale — Amazon (AWS), Microsoft (Azure), Google, and Meta are the most commonly cited examples. They build and run their own vast networks of data centers.

Neocloud
A newer breed of company that builds data centers and rents out raw GPU computing power, without offering the full suite of cloud software services a hyperscaler does. They've emerged specifically to meet AI-driven GPU demand (examples often cited include CoreWeave and Lambda).

Colocation (“Colo”)
A facility where multiple different companies rent space, power, and cooling for their own servers inside a shared building, rather than each building their own data center from scratch.

Cloud Computing
Renting computing power, storage, or software over the internet from someone else's data center, instead of buying and running your own hardware.

On-Premises (“On-Prem”)
The opposite of cloud computing: a company owns and runs its computers in its own building rather than renting capacity elsewhere.

Data Center Campus
A cluster of multiple large data center buildings on one site, often the scale being described when a company announces a multi-billion-dollar, multi-gigawatt AI facility.

Tier Rating (e.g. Tier III, Tier IV)
An industry classification (I through IV) describing how much redundancy and fault tolerance a data center has — Tier IV facilities are built so that even planned maintenance or equipment failure shouldn't cause downtime.

Uptime
The percentage of time a data center (or the services running in it) is actually operational, as opposed to down for maintenance or an outage. Even “99.99% uptime” still allows for roughly 53 minutes of downtime a year.

Edge Data Center
A smaller facility placed closer to the people or devices using it, to reduce latency — as opposed to a giant, centralized campus that may be hundreds of miles from its users.

Sovereign AI / Sovereign Cloud
The push by individual countries to build or control their own AI infrastructure and data centers domestically, rather than depending entirely on foreign-owned cloud providers.

AI Factory
A marketing and industry term (popularized by Nvidia) for a data center purpose-built around AI computation, framed less like a traditional IT facility and more like an industrial plant that manufactures “intelligence” as its output.

Environment & Sustainability
Terms that come up around the resource footprint of AI data centers.

Carbon-Free Energy
Electricity generated without direct carbon emissions — including renewables like wind and solar, as well as nuclear power. Many operators have made public pledges to run on carbon-free power.

Renewable Energy Credit (REC)
A certificate representing one unit of renewable electricity generated somewhere on the grid. Companies buy RECs to offset their electricity use on paper, even if the actual power reaching their data center isn't directly renewable.

Small Modular Reactor (SMR)
A newer, smaller class of nuclear reactor designed to be manufactured in factories and deployed faster and more cheaply than traditional nuclear plants — several tech companies have signed deals to use SMRs to power future AI data centers.

Water Consumption
The actual volume of water used for cooling, which can be substantial for large facilities using evaporative cooling — a source of local community concern in water-stressed areas.

Compiled for a general, non-technical reader. Terminology in this space evolves quickly as the industry grows — treat this as a starting reference rather than an exhaustive or permanently up-to-date list.