AI Data Centre FAQ & Glossary

1. What is the difference between a cloud data centre and an AI data centre?
Cloud data centres provide general-purpose computing services — storage, databases, web hosting — using CPU-based servers. AI data centres are optimized for GPU/TPU workloads, with much higher power density, specialized cooling, and high-speed networking. Many hyperscalers are building AI-specific wings within or adjacent to their cloud campuses.

2. How much does it cost to build an AI data centre?
Construction costs range from $10 million to $25 million per MW of IT load, depending on location, cooling method, and power infrastructure. A 100 MW AI data centre can cost $1–2.5 billion to build before hardware is installed. Hardware (GPUs, networking) adds several billion more for a large-scale facility (CBRE Research, 2025).

3. How much electricity does an AI data centre use?
Large AI data centres range from 100 MW to 500 MW of power consumption. A 100 MW facility running continuously uses approximately 876 GWh of electricity per year — roughly equivalent to the annual electricity consumption of 80,000 U.S. homes (U.S. EIA, 2024).

4. Which companies operate the largest AI data centres?
Microsoft, Google, Amazon, Meta, and Oracle operate the largest AI-focused facilities. Specialized AI cloud providers like CoreWeave, Lambda Labs, and Together AI also operate significant GPU clusters. In China, Alibaba Cloud, Huawei, and ByteDance lead.

5. Are AI data centres bad for the environment?
The environmental impact is significant and growing. The IEA estimated global data centre electricity consumption at 460 TWh in 2022, with projections exceeding 1,000 TWh by 2026. While renewable energy procurement has increased, absolute carbon emissions from data centres are rising in most regions due to the scale of new builds.

6. What is PUE and why does it matter?
PUE stands for Power Usage Effectiveness. It is the ratio of total facility energy to the energy used by IT equipment. A PUE of 1.0 is perfect efficiency. Lower is better. AI data centres typically target PUE of 1.1–1.3. Google's global fleet averaged 1.10 in 2023 (Google Environmental Report, 2024).

7. What is liquid cooling and why is it used in AI data centres?
Liquid cooling involves circulating water or engineered fluid near or directly on processors to remove heat. GPU clusters generate 30–120 kW per rack — far too much heat for air cooling to handle efficiently. Liquid cooling is faster, more efficient, and allows higher rack density.

8. What is sovereign AI infrastructure?
Sovereign AI infrastructure refers to AI compute capacity built and controlled within a country's borders, subject to its laws. Governments pursue it to protect data privacy, avoid reliance on foreign cloud providers, and maintain strategic autonomy in AI development.

9. What is a hyperscaler?
A hyperscaler is a company that operates data centres at an exceptionally large scale, providing cloud services globally. The major hyperscalers are Amazon Web Services, Microsoft Azure, Google Cloud, Alibaba Cloud, and Meta. They are the primary builders and operators of AI data centres.

10. Can AI data centres use nuclear power?
Yes, and they increasingly are. Microsoft signed a deal to restart the Crane Clean Energy Centre (formerly Three Mile Island Unit 1) in Pennsylvania, providing nuclear power directly to its data centres. Multiple SMR developers have signed agreements with data centre operators for future nuclear power supply.

11. How long does it take to build an AI data centre?

Greenfield construction typically takes 2–4 years from land acquisition to first power-on, primarily due to permitting, grid interconnection timelines, and construction. Modular or prefabricated builds can be faster — some operators have deployed containerized GPU clusters in under 12 months.

12. What is inference vs. training in AI computing?
Training is the computationally intensive process of building an AI model from data. Inference is using an already-trained model to produce outputs. Training requires the most powerful GPU clusters for sustained periods. Inference requires high throughput and low latency, and as AI adoption scales, inference is becoming the dominant workload by volume.

13. What role do data centre REITs play?
Real Estate Investment Trusts (REITs) like Equinix, Digital Realty, and Iron Mountain own and operate colocation data centres, leasing space and power to tenants. As AI demand surged, data centre REITs became among the best-performing REITs in 2023–2024. They increasingly offer AI-ready, high-density colocation services.

14. What is immersion cooling?
Immersion cooling submerges entire servers in a thermally conductive but electrically inert fluid (such as engineered fluids from 3M or Engineered Fluids). The fluid absorbs heat and is circulated to an external heat exchanger. It allows rack densities exceeding 100 kW and eliminates the need for traditional air cooling infrastructure.

Glossary

- **AI Data Centre:** A facility purpose-built to host GPU or TPU clusters for training and running AI models, with high-density power, liquid cooling, and ultra-fast networking.
- **GPU (Graphics Processing Unit):** A processor designed for parallel computation. Originally for graphics rendering, now the primary chip for AI training. NVIDIA dominates this market.
- **TPU (Tensor Processing Unit):** Google's custom AI chip, designed specifically for neural network computations. Used internally by Google and available on Google Cloud.
- **Hyperscaler:** A company that operates cloud infrastructure at massive global scale — primarily AWS, Microsoft Azure, Google Cloud, Meta, and Alibaba Cloud.
- **PUE (Power Usage Effectiveness):** A metric for data centre energy efficiency. PUE = Total facility power ÷ IT equipment power. Lower is better; 1.0 is perfect.
- **Inference:** Using a trained AI model to generate outputs (answers, images, predictions). Distinct from training, which builds the model.
- **Training:** The process of building an AI model by processing large datasets and adjusting model parameters. The most compute-intensive AI workload.
- **Liquid Cooling / Direct Liquid Cooling (DLC):** Cooling method in which water or fluid is piped directly to server components to remove heat. Required for high-density AI racks.
- **Immersion Cooling:** Submerging servers in dielectric fluid for cooling. Allows very high rack density (100 kW+) with minimal water consumption.
- **Sovereign AI:** AI compute infrastructure built and operated within a country's borders, under its legal and regulatory control.
- **Colocation (Colo):** A data centre where multiple tenants rent space, power, and connectivity — as opposed to a single-tenant hyperscaler campus.
- **InfiniBand:** A high-speed networking standard used to connect GPU clusters within AI data centres, offering very low latency and high bandwidth.
- **SMR (Small Modular Reactor):** A compact nuclear reactor (typically under 300 MW) that can be factory-built and deployed faster than traditional nuclear plants. Seen as a future power source for AI data centres.
- **NVLink / NVSwitch:** NVIDIA's proprietary interconnect technologies allowing multiple GPUs to communicate at very high speeds within a server or across a server cluster.
- **Power Purchase Agreement (PPA):** A long-term contract between an electricity generator (e.g., a wind farm) and a buyer (e.g., a data centre operator) to purchase power at a fixed price.