



White Paper on AI Digital Infrastructure (AIDI)

February 2025

Executive Summary

The growing integration of Artificial Intelligence (AI) technologies into business operations, research, and public services has highlighted the necessity of advanced digital infrastructure, especially data centres, to support the computational and data storage needs of AI applications. AI models, particularly those using deep learning, require immense amounts of data, computational power, and real-time processing capabilities. This white paper explores the vital role of AI digital infrastructure in data centres, examining their design, components, key challenges, and the strategies needed to create an optimised environment for AI workloads. We also discuss the emerging trends in data centre technology, such as edge computing, liquid cooling, and the increasing use of GPUs and TPUs, to meet the needs of next-generation AI.



Table of Contents

1. Introduction

2. The Role of Data Centres in AI Digital Infrastructure

- 2.1 Computational Power
- 2.2 Storage and Data Management
- 2.3 Connectivity

3. AI-Specific Requirements for Data Centres

- 3.1 GPU and TPU Integration
- 3.2 Energy Efficiency and Sustainability
- 3.3 Latency and Bandwidth Demands
- 3.4 Power and Cooling

4. Emerging Trends in AI Data Centres

- 4.1 Edge Computing and Edge AI
- 4.2 Liquid Cooling and Green Data Centres
- 4.3 Software-Defined Infrastructure and Automation

5. Key Challenges in AI Data Centre Infrastructure

- 5.1 Scalability and Flexibility
- 5.2 Data Security and Privacy
- 5.3 Cost Management

6. Best Practices for Optimising AI Data Centres

- 6.1 Infrastructure Planning and Design
- 6.2 Energy-Efficient Practices
- 6.3 Security and Compliance Frameworks

7. Case Studies and Industry Applications

- 7.1 AI in Healthcare
- 7.2 AI in Finance and Autonomous Systems
- 7.3 Smart Cities and IoT Integration

8. Conclusion



Introduction

1. Introduction

Artificial Intelligence (AI) is transforming industries worldwide, from healthcare to finance to logistics. However, AI's power and effectiveness rely heavily on robust digital infrastructure, especially in data centres. The computing demands for AI applications are immense, as AI models, particularly deep learning networks, require vast amounts of computational power, high-performance storage, and efficient data processing capabilities. Data centres are the backbone of AI infrastructure, providing a dedicated environment to process the massive datasets AI systems rely on.

In this white paper, we explore the components of AI digital infrastructure in data centres, analyse the challenges and solutions for optimising these environments, and look at emerging trends that are shaping the future of AI infrastructure.





The Role of Data Centres in AI Digital Infrastructure

2. The Role of Data Centres in AI Digital Infrastructure

Data centres serve as the heart of AI operations, providing the physical space and resources to host AI models and process large datasets. The main functions of AI-specific data centres are:

2.1 Computational Power

AI workloads, particularly those based on machine learning (ML) and deep learning (DL), require substantial computational resources. Traditional data centres with CPU-only infrastructure are often insufficient for these tasks. Instead, AI workloads demand specialised hardware, including:

- **Graphics Processing Units (GPUs):** GPUs are designed to handle parallel processing tasks, making them ideal for training AI models, especially deep neural networks (DNNs). Their ability to process multiple data streams simultaneously significantly reduces the time needed for training AI models.
- **Tensor Processing Units (TPUs):** Developed by Google, TPUs are custom hardware accelerators optimised for AI workloads, particularly for machine learning tasks. They offer even more specialised performance than GPUs, enabling faster processing for AI applications.

2.2 Storage and Data Management

Data management is another critical aspect of AI infrastructure. The large amounts of data AI models need for training and validation necessitate high-performance storage solutions. These storage systems must be able to handle structured and unstructured data types and support fast retrieval for real-time AI applications. Key aspects of AI data centre storage include:

- **High-Capacity Storage Systems:** Data lakes and distributed storage systems are crucial for storing massive volumes of data. These systems must support high throughput and quick access for processing large datasets.
- **Data Tiering:** AI systems benefit from data tiering, which optimises the management of data by storing frequently accessed data in high-performance storage (e.g., SSDs) while less critical data can be stored in lower-cost solutions.

2.3 Connectivity

The success of AI applications is closely linked to the performance of networking systems in the data centre. AI workloads, particularly those involving real-time processing or collaboration across distributed systems, require high bandwidth and low-latency connections. This is especially true for applications such as autonomous vehicles, industrial automation, and smart city services. The key factors for AI data centre connectivity include:

- **High-Bandwidth Networks:** Data centres must support large-scale data transfer between GPUs/TPUs and storage systems. High-speed connections such as InfiniBand and 100-900Gbps Ethernet are critical.
- **Low-Latency Connections:** Reducing latency is essential for AI applications like real-time decision-making and streaming analytics. Optimising networking paths to minimise delay is crucial in data centre design.



AI-Specific Requirements for Data Centres



3. AI-Specific Requirements for Data Centres

AI workloads have unique requirements that differ from traditional IT operations. To meet these demands, data centres must be designed with the following considerations:

3.1 GPU and TPU Integration

The growing importance of GPUs and TPUs in AI applications requires that data centres are optimised for the deployment and management of these processors. Unlike CPUs, GPUs and TPUs are more efficient at handling parallel tasks, which is vital for AI tasks like training neural networks.

Data centres must offer specialised racks and power systems that support the unique power and cooling demands of these high-performance accelerators. Managing these processors effectively requires a deep integration between hardware and software, including virtualisation platforms designed to optimise AI workloads.

3.2 Energy Efficiency and Sustainability

AI workloads, particularly those related to deep learning, consume a significant amount of energy. As AI becomes more widespread, the demand for computational resources will only increase, putting pressure on energy consumption. Data centres must be designed with energy efficiency in mind, focusing on:

- **Efficient Cooling Systems:** Advanced cooling techniques such as liquid cooling and immersion cooling can help reduce the energy required to maintain optimal temperatures for GPUs and TPUs.
- **Renewable Energy Sources:** Many companies are transitioning to using renewable energy to power data centres, aiming to minimise the environmental impact and reduce operational costs.

3.3 Latency and Bandwidth Demands

AI applications that require real-time decision-making, such as autonomous vehicles or industrial robotics, demand low-latency communication and high-bandwidth connectivity. Traditional data centres may not be sufficient to meet these needs. Emerging technologies such as edge computing can help address this challenge by processing data closer to the source, reducing the need to transmit large volumes of data to centralised data centres.

3.4 Power and Cooling

AI workloads running on the latest NVIDIA GPU's such as the A100, H100, H200 or the GB200, require an immense amount of power and cooling. Due to the high power draws of cutting edge NVIDIA GPU's, cooling infrastructure, not commonly seen in legacy data centres, is required to cool the GPU's and prevent overheating.

A legacy data centre built in the 2000's is quite often designed around a power density of 3kW per cabinet or 1.8kW/m² running DX cooling with CRACs, whereas GPU infrastructure typically requiring a rack density of up to 50kW (30kW/m²) air cooled utilising Rear Door Heat Exchangers or up to 200kW (120kW/m²) with Direct to Chip liquid cooling such as that seen on the NVIDIA GB200. These more advanced cooling technologies required a Chilled Water or Adiabatic Cooling System with a primary and secondary chilled water loop system and CDU.

GPU clusters will also require multiple 3-phase power supplies to power the high density equipment.



Emerging Trends in AI Data Centres

4. Emerging Trends in AI Data Centres

As AI continues to evolve, new trends and technologies are shaping the development of data centres. These include:

4.1 Edge Computing and Edge AI

Edge computing refers to processing data closer to the source (e.g., IoT devices, sensors) rather than sending it to a centralised cloud or data centre. This reduces latency and bandwidth requirements. Edge AI, where AI models are deployed directly on edge devices or local servers, enables real-time decision-making without relying on distant data centres.

This trend is particularly relevant in applications like autonomous vehicles, robotics, and industrial IoT, where low latency and real-time processing are critical.

4.2 Liquid Cooling and Green Data Centres

AI data centres require advanced cooling systems to manage the heat generated by high-performance processors. Liquid cooling technology has emerged as a more energy-efficient alternative to traditional air cooling methods, enabling higher density and lower energy consumption.

Green data centres, which focus on sustainability through energy-efficient practices and renewable energy sourcing, are becoming a priority in AI infrastructure development.

4.3 Software-Defined Infrastructure and Automation

Software-defined infrastructure (SDI) and automation technologies allow for greater flexibility and efficiency in managing AI workloads. SDI enables the abstraction of hardware resources, allowing for the dynamic allocation of resources based on workload demands. Automation can help optimise data centre operations, ensuring efficient provisioning, monitoring, and management of AI workloads.





Key Challenges in AI Data Centre Infrastructure

5. Key Challenges in AI Data Centre Infrastructure

Despite the growing importance of AI, there are several challenges when developing and maintaining AI-specific data centre infrastructure:

5.1 Scalability and Flexibility

AI workloads are expected to grow rapidly, with increasingly large datasets and more complex models. Data centres must be scalable to accommodate this growth, requiring flexible hardware and software configurations that can evolve with changing needs.

5.2 Data Security and Privacy

As AI systems handle sensitive data, security and privacy become paramount. Ensuring compliance with data protection regulations, such as GDPR, is essential for maintaining customer trust and avoiding legal challenges.

5.3 Cost Management

The cost of building and maintaining AI-optimised data centres can be prohibitive. Managing the costs of specialised hardware, energy consumption, and maintenance is a significant challenge for organisations.





Best Practices for Optimising AI Data Centres

6. Best Practices for Optimising AI Data Centres

To optimise AI data centres, organisations should adopt the following best practices:

Infrastructure Planning and Design

Invest in specialised hardware and infrastructure that can support the unique demands of AI workloads, including GPUs, TPUs, and high-speed networking.

Energy-Efficient Practices

Implement liquid cooling, renewable energy sources, and other green technologies to reduce energy consumption and environmental impact.

Security and Compliance

Adhere to data protection laws and use encryption, access controls, and other security measures to ensure the protection of sensitive data.





Case Studies and Industry Applications

7. Case Studies and Industry Applications

Real life use cases for the application of AI.

7.1 AI in Healthcare

AI has transformative potential in healthcare, including applications such as medical imaging, drug discovery, and patient monitoring. Data centres enable the processing of large-scale patient data and the deployment of AI models that assist healthcare professionals in diagnosing and treating patients.

7.2 AI in Finance and Autonomous Systems

Financial institutions rely on AI for fraud detection, risk modelling, and algorithmic trading. AI data centres provide the computational power necessary to process real-time data streams, making financial transactions more secure and efficient.

7.3 Smart Cities and IoT Integration

AI-driven smart city initiatives, such as intelligent traffic management and environmental monitoring, require robust data infrastructure to process the vast amounts of data generated by IoT devices.





Conclusion

8. Conclusion

AI is revolutionising industries worldwide, and data centres play a critical role in enabling the deployment of AI technologies. The computational power, storage, connectivity, and security provided by data centres are essential for supporting AI workloads. As AI technology continues to evolve, data centres must adapt to meet new demands, including energy efficiency, low-latency processing, and the integration of advanced hardware like GPUs and TPUs. By investing in scalable, secure, and sustainable data centre infrastructure, businesses can unlock the full potential of AI while ensuring long-term growth and success.





Lunar Digital is the trading name of Lunar Digital Solutions Limited.
Kilburn House, Lloyd Street North, Manchester, M15 6SE Tel: +44(0)330 555 1550
Registered in England and Wales number 13872813