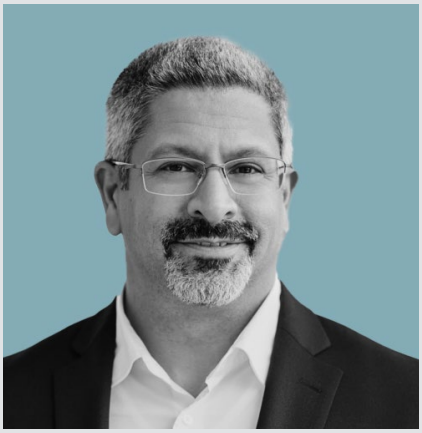
DTCP

Marketing Communication

The Impact of Artificial Intelligence on the Data Center Market



Zahl Limbuwala

Operating Partner, DTCP Infra

In this short paper, Zahl gives a brief outline of the development of artificial intelligence (AI) from its niche existence in the 1950s to the success story of large language models. The evolution of artificial intelligence has far-reaching implications for data centers, including significant alterations in their design, driven by requirements on power density and interconnectivity. This changing market offers excellent investment opportunities.

About the Author

Zahl Limbuwala is an Operating Partner at DTCP, focusing on data centers, and Chairman of the Supervisory Board at maincubes, a colocation data center provider. He has more than thirty years of industry experience, most recently as Chief Operating Officer of Atlas Edge, one of Europe's key operators of edge data center infrastructure. Prior to that, he was responsible for strategy and strategic client engagement at CBRE's data center facilities management business unit, which managed over five hundred data centers globally. Zahl has led technology organizations across software development, IT hardware, communication networks, and data center facilities.

Brief Outline of the Development of Artificial Intelligence

How has AI evolved since its invention and what has fueled its success story?

A. Since psychologist Frank Rosenblatt created the first artificial neural network in 1956, AI has seen an impressive evolution. What has really stepped us into the era of AI is the exponential growth of raw compute power and, more recently, the breakthrough by Google in the creation of the 'Transformer'. This is an approach to neural network training that underpinned the explosion of large language models (LLMs). **Application-Specific Integrated Circuits (ASICs) such as Google's tensor processing units (TPUs) and Nvidia's graphics processing units (GPUs) have given us 'super-computing' at a scale, able to handle and process neural networks with billions and even trillions of parameters.** The parameters are the weightings on the nodes in the neural network.

How have these 'super-computing' enablers, the TPUs and GPUs, contributed to the acceleration of AI research and applications?

A. For decades, the Central Processing Unit (CPU) was developed to produce general-purpose computational power at an ever-lower cost per unit compute. The CPU became highly commoditized, resulting in the domination of companies such as Intel and AMD. There have always been specialized processing units, but these were used in niche and embedded, mostly industrial applications. The emergence of vector graphics, computer-generated images, and other computationally intense applications created the first processors designed and produced at scale, that were optimized for specific mathematical functions, such as scalar, vector, matrix, tensor mathematical functions. These are exactly the functions required to build and compute neural networks at scale and make computing times and costs "reasonable".

With neural networks being the backbone, how have deep learning architectures developed and what role do they play in modern AI systems?

A. The field of machine learning has exploded with the enablement of massive-scale application-specific computing but also breakthrough developments in neural network training techniques. There are many approaches being developed today each with their characteristics, strengths and weaknesses. For example, one well-publicized weakness of LLMs is 'hallucination' which is the name given when responses to questions posed to a model are factually incorrect, but the model is convinced of their correctness.

Interestingly, these machine learning breakthroughs have led to significant investments in AI foundation models, hardware infrastructure deployment for both model training and inference – as the application phase is called – and consequently a net-new demand for data center capacity. **This new demand comes on-top of the already expanding need for cloud capacity.** In Europe in particular, current demand is being driven by data sovereignty considerations as well as the net increase of end users.

Artificial Intelligence is driving Structural Change in the Data Center Market

How has the widespread adoption of AI impacted the scalability requirements of data centers?

A. The development of, investment in and adoption of AI has led to a massive increase in demand for data center capacity. **The aggregate predicted growth numbers are very large – for example, some market participants forecast even need to get to 1GW campus developments and single leases of 300 MW+.**

However, not all demand is equal:

- The training of AI models typically requires a large amount of centralized computing. Centralized because these training systems can be thought of as a form of 'super-computer' and thus the connectivity between all the nodes in the computer needs to be low latency and high bandwidth. This results in larger and larger single building data centers but also larger and larger campuses with high speed/high bandwidth intra-campus connectivity.
- As some business users will be using some of their most sensitive data within the AI model, data security and data sovereignty at the sites of AI inference - the models used by end users - will be of great importance.
- Given the highly specialized nature of the computer hardware (GPU, TPU, etc.) their power requirements are higher than commodity CPUs thus dealing with the waste heat from these chips requires direct-to-chip (or immersed) liquid cooling. **Overall, the impact on data centers is that a higher power density needs to be considered in their design.**

What specific computational resources of GPUs, TPUs and CPUs, are in high demand due to AI workloads?

A. Almost any ASIC is in demand, either directly as a hardware component to procure and install in an operator's data center or internally to make available to a customer base via some sort of rental/leased capacity model.

Do data centers need to adapt their infrastructure to meet the special requirements of AI applications?

A. Most data center designs being created to accommodate AI workloads are considering the following requirements:
- Secure as much available grid power as possible for the location
- Consider on-site/renewable options in addition to grid
- Design for increased power density, given the continuous upward trend; some project that IT rack power could reach 30kW by the decade's end[1]
- Include a cooling system with closed water circuit (doesn't consume water) in order to accommodate direct-to-chip and immersion-based cooling systems

1 Uptime Institute: Global Data Center Survey 2023, July 2023, Five Data Center Predictions for 2023, January 2023

Besides power and cooling, what are the specific design considerations for AI workloads in terms of high-speed connectivity or specialized storage solutions?

A. Another component for AI training, in particular, is the design accommodating the need for a significant increase in networking between compute nodes. Within the data center building, around a data center campus and then externally for ingress and egress data transfer.

The last key point in data center design is the storage architecture. How has this evolved to cope with the growing volumes of data generated by AI applications?

A. Data storage has evolved in recent years from electro-mechanical systems, Hard Disk Drives (HDDs), that use a spinning metal disk on which data is stored magnetically, to now Solid State Disks (SDDs) which are entirely electrical (Integrated Circuits) and have no mechanical or moving parts. Optical technologies will become increasingly important in the future. For example, Microsoft's Project Silica leverages discoveries in ultrafast laser optics and AI and is able to store data in quartz glass. **While these inventions are not yet ready for large-scale adoption, they highlight the documented need to ensure higher data storage density with little or no data loss over time.**

Conclusion

Today, data centers are acknowledged as an alternative asset class, positioned between real estate and infrastructure. The global demand is propelled by the ongoing digitalization of global economies and societal change.
The utilization of diverse digital services, facilitated by advancing technologies, has become feasible due to the consistent and predictable growth of computing power. The emergence of machine learning and AI introduces new possibilities, placing additional demands on the digital infrastructure of data centers and communication networks.

6.1%

is the expected compound annual growth rate (CAGR 2023-2028) for revenue in the data center market, resulting in a forecasted market size of **€400 billion** in 2028[2]

DTCP Infra, comprising 14 individuals with strong investment backgrounds, possesses a solid track record and deep knowledge within the field. With full value chain competency, we stand as entrepreneurial specialists. Notably, our senior team members boast a 15+ years track record, marked by about 100 transactions in the digital infrastructure sector, affirming our expertise and proven success in navigating this complex landscape. The team is complemented by operational partners who provide expertise in efficiently managing and optimizing the day-to-day aspects of digital infrastructure projects.

The joint efforts underpin our holistic approach to ensure sustainable progress across the entire lifecycle of our portfolio companies.

2 Statista Market Insights 2023, Global Data Center Investor Sentiment Survey 2022