IRON MOUNTAIN®
DATA CENTERS

Planning your AI-ready infrastructure

Top 5 factors to consider

IRON MOUNTAIN®
DATA CENTERS

Planning your AI-ready infrastructure

Top 5 factors to consider

AI: a world of opportunities and challenges

It is now over two years since Chat GPT was launched and, with new AI-focused funds being launched continually, investment shows no signs of slowing down. AI systems are trained on large amounts of data to identify patterns. Generative AI goes a step further, using complex systems and models to generate new outputs in the form of images, text, or audio using natural language prompts.

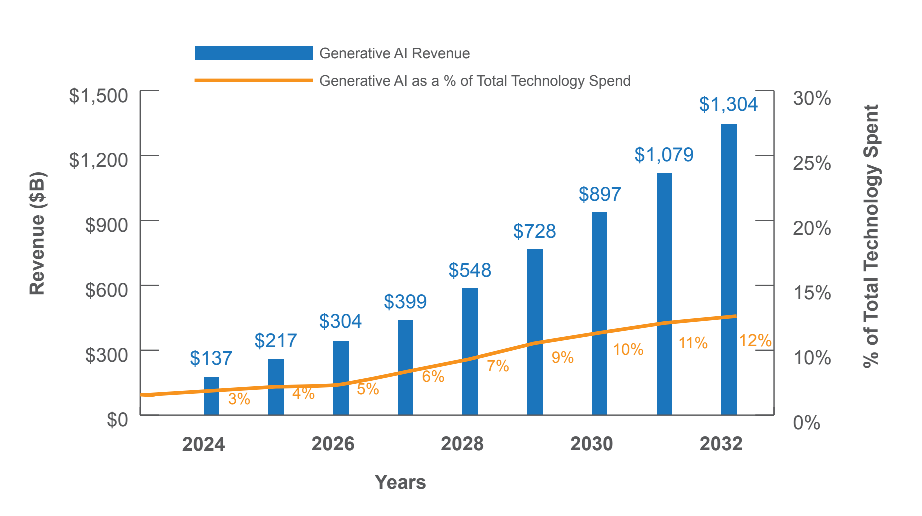
This new model promises massive social and economic value as it can handle multiple repetitive tasks quickly and accurately, and break down communication barriers between humans and machines.

According to IDC, AI could have a cumulative global economic impact of $19.9 trillion through 2030 and drive 3.5% of global GDP by 2030. Hundreds of billions of dollars are being invested in the components and physical infrastructure that will run AI.

However, not all of today's data centers are AI-ready. Power and cooling configurations, facility architecture, site selection, connectivity and environmental impact are all critical to success.

This e-book sets out the top five factors that businesses should consider when planning their High-Performance Computing (HPC) or AI infrastructure needs:

1. Primary & backup power
2. Flexibility & scalability
3. Cooling considerations
4. Connectivity & latency
5. Climate impact



In one of the more conservative forecasts, Bloomberg estimates that generative AI revenue will exceed $1 trillion by 2031, accounting for over 10% of total technology spend. The AI revolution will be physically enabled in the data center and will create a step change in the infrastructure market.

1 Primary & backup power

Generative AI models use graphics processing unit (GPU) chips which require 10–15 times the energy of a traditional CPU, because they have more transistors in the arithmetic logic units. This requires more compute, network, and storage infrastructure than any previous technology. Many generative AI models have billions of parameters and require fast and efficient data pipelines in their training phase, which can take months to complete, and these training phases then need to be repeated as new data is added.

Large Language Model (LLM): Training & Power

ChatGPT 3.5 had 175 billion model parameters and was trained on over 500 billion words of text. According to one article, this class of generative AI model requires a ten to a hundred-fold increase in computing power to train models compared to the previous generation. This doubles overall demand every six months. To train a ChatGPT 3.5 model requires 300-500 MW of power.

After the training phase, the second phase in generative AI delivery is inference, or using the application to respond to inquiries and return data results. Power consumption during inference is still very high: A Google search can power a 100w lightbulb for 11 seconds, but a single ChatGPT session consumes 50-100 times more power than this.

Generative AI training has to take place in specialized HPC data centers. Colocation providers are scaling up their power for new builds in anticipation of AI requirements, and implementing some key structural changes to meet this demand. Key features to look for in the data center include high-density power supplies to servers that can provide 30-100 kW/rack or higher as opposed to the traditional 5-10 kw/rack. They should also have robust backup power for 24/7 operation of high-powered infrastructure.

The new requirements of AI are driving innovation in low carbon solutions such as battery, hydrogen, and nuclear power to ensure 24/7 uptime for power-hungry applications.

2 Flexibility & scalability

Ramping up power usage requires new approaches to design and construction. Modular design is better equipped to cope with extra complexity and new datasets as AI and HPC arrays expand. A modular approach enables faster scalability and new cooling methods, as well as being quicker and cheaper to scale up. One report estimates that modular data centers are at least 60% faster to deploy and provide 13% or more cost savings than traditional data center power and cooling infrastructure.

AI in action: exponential growth

One of IMDC's healthcare customers has developed a supercomputer for AI-driven imaging apps. While the total consumption is not massive and training cycles are much shorter than for LLMs, the data growth curve of this supercomputer has been steep. It began in 2018 with just 10-50,000 images and achieved 85% accuracy. Now it uses up to half a billion images with accuracy of 95% and runs 50,000 deep learning training experiments per month. Despite the compactness and efficiency of the GPUs, a few racks in the data center have become a full module of 60 racks with 26 petabytes of processing power storing close to 2 billion datasets. This year a petaflop of processing power will be needed.

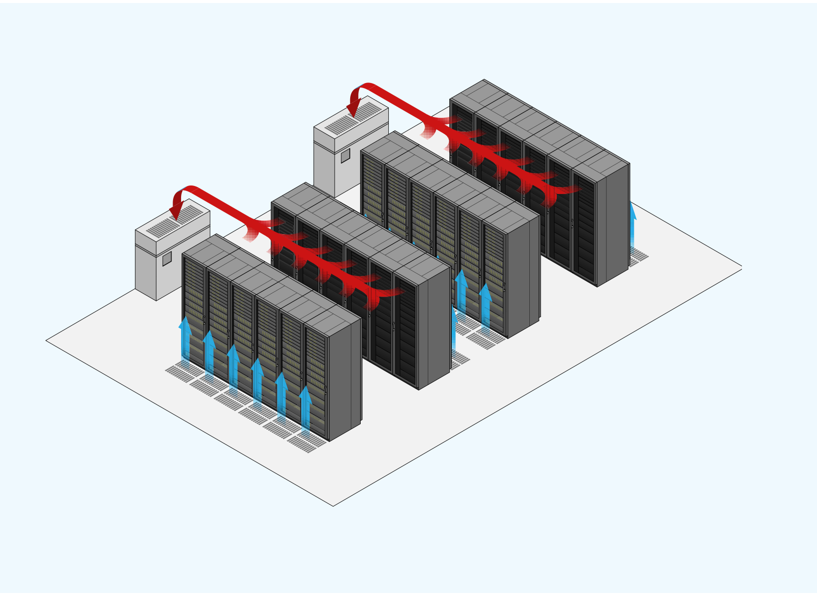
3 Cooling considerations

Some computing hardware now enables power densities exceeding 100 kW/rack and the peak density in the data center could reach 150 kW/rack over the next few years. This increasing compression of computing power generates enormous heat which presents new challenges for data center cooling infrastructure.

The density of an AI or HPC deployment determines its unique cooling needs and air can only absorb a certain amount of heat. Deep learning and Generative AI require alternative solutions due to their complexity, and these will need specialized infrastructure such as direct liquid cooling (DLC), air-assistant liquid cooling (AALC), and/or a rear-door heat exchange. With liquid cooling, water or another contact fluid flows through the racks. Many experts consider this a key enabling technology for AI in data centers.

Not all AI or HPC deployments will require liquid cooling. Inferencing tends to be less power-intensive than training and may be able to be cooled with traditional air-cooling techniques. ML (Machine Learning) also requires fewer resources.

Operating the mixed densities customers require to run AI requires a more sophisticated approach to cooling. Not every data center provider will have the specialized knowledge or flexible infrastructure needed to enable this.



Liquid Cooling Benefits

Once liquid cooling has been implemented, there are benefits. More usable space can become available, because the technology takes up less space. Also, because it operates in a closed circuit, liquid cooling can lend itself to heat exchange. With Direct Liquid Cooling, the temperature of the waste heat rises, making it more worthwhile to reuse, and heat transfer between the data center and district heating becomes more energy-efficient than air cooling. It is also a great deal quieter to run.

4 Connectivity & latency

As new GPU chip designs emerge using 3D packaging and fiber-optic rather than copper connections, it is clear that rich low-latency data connectivity is the key to the future speed and efficiency of AI. The same principles apply at the macro or data center level. Data centers require high-bandwidth network connectivity to facilitate the efficient exchange of the large datasets necessary for training AI models. Advanced networking technologies such as high-speed optical fiber and low-latency inter-connected networks play a vital role in ensuring efficient data transfer and communication between servers. Modern AI applications require high-bandwidth, lossless, low-latency, scalable networks that can interconnect hundreds and thousands of GPUs at speeds as high as 400 or 800 Gbps.

Latency variations

Infrastructure for training models is generally less location-sensitive than for traditional consumer clouds or for inference. Locations with plentiful low-carbon energy are therefore more popular, given the constraints around power in more developed markets. Inference requires more geographically dispersed infrastructure that can scale quickly and provide connectivity to the applications with lower latency, particularly for real-time or immersive applications. This means more data centers in more locations, a model which differs from the centralized public cloud model that currently supports most applications.

Less urban locations with plenty of renewable power and space can be suitable for AI training, reducing environmental impact by using low-carbon energy like geothermal or hydro power. Our innovative 'run of river' hydro purchase agreement with Rye Development, for instance, will ensure that generators are added to dams in Pennsylvania and West Virginia. These have the potential to create up to 150 MW of 24/7/365 customer power over the next ten years.

5 Climate impact

Is an AI-driven world on a collision course with emission reduction targets? Not if users prioritize sustainability in their planning. Major commitments have been made by industry leaders to efficient use of resources like water and power, plus elimination or offsetting of emissions. All of these targets are fact-based and non-negotiable. As AI deployments use so much power both providers and users need to address these issues fully and transparently at the outset.

Energy Efficiency

Optimizing efficiency is the baseline for sustainable operations. Chip redesign has the potential to improve compute efficiency hugely, and data center operators need to match these efficiencies, maximising PUE.

AI has a key role to play in this process due to its ability to process huge amounts of data. Digital twinning, for instance, has shown a lot of potential already. IMDC piloted the DCverse AI-Powered Cognitive Digital Twin in its Singapore facility. It uses real-time data collection, advanced physics-informed machine learning algorithms, and predictive analytics to deliver optimized control policy recommendations for cooling, and delivered a PUE improvement of 3.5% within three months. This solution has the potential to deliver similar levels of improvement at other IMDC facilities and more widely across the data center industry.

Renewable Power

AI makes the use of low-carbon energy a critical consideration. Most hyperscalers and a growing number of colocation providers have been growing the green grid and eliminating carbon. Data center operators are already the biggest buyers of renewables in the world, and the more sustainable colocation providers are up there with them. Iron Mountain is now one of the world's top 20 renewable buyers.

Complete decarbonization

Some data center businesses are now going a step further, pursuing complete decarbonization of power through hour by hour matching of site consumption with local carbon free generation.

While we work to extend 24/7 carbon-free energy to all of our facilities and customers we also offset 100% of our power with renewables, something we have been doing since 2017.

Planning your AI infrastructure needs

Huge advances are being made in data center design and operations to accommodate AI and HPC configurations, but not all facilities or operators can fully support the new technology. New requirements will inevitably emerge as tailored AI applications are launched to specific sectors and markets. In the meantime, if you are considering an AI deployment bear the following key factors in mind:

Power

Can your provider offer sufficient power in the right high-density configurations? Look for high-density power supplies to servers that can provide 30-100 kW/rack or higher, with robust low-carbon backup power.

Flexibility & scalability

Facilities need to be both modular and scalable. Your processing power and storage requirements are likely to rise exponentially

Cooling

Are all of the advanced cooling solutions in place to cater for all of your needs, from air cooling to DLC? Remember you are likely to require a mix of densities and this may require a mix of cooling technologies.

Connectivity & latency

Can you access the sort of scalable bandwidths you need for training and the latencies you need for inference?

Climate impact

Efficiency needs to be optimized, and AI can help with this. Also all energy (and backup) sources should be as close to carbon-free as possible, preferably taken from the same grid.

If you are interested in finding out more about planning, locating and activating your AI-ready data center infrastructure, please get in touch:

NL: +31 800 272 4433
UK: +44 844 417 8379
DE: +49 800 408 0000
US: 833.IRM.colo
Contact Us

IRON
MOUNTAIN®
DATA CENTERS