IRON MOUNTAIN®
DATA CENTERS

Customer Solution Brief

Generative AI

Contents

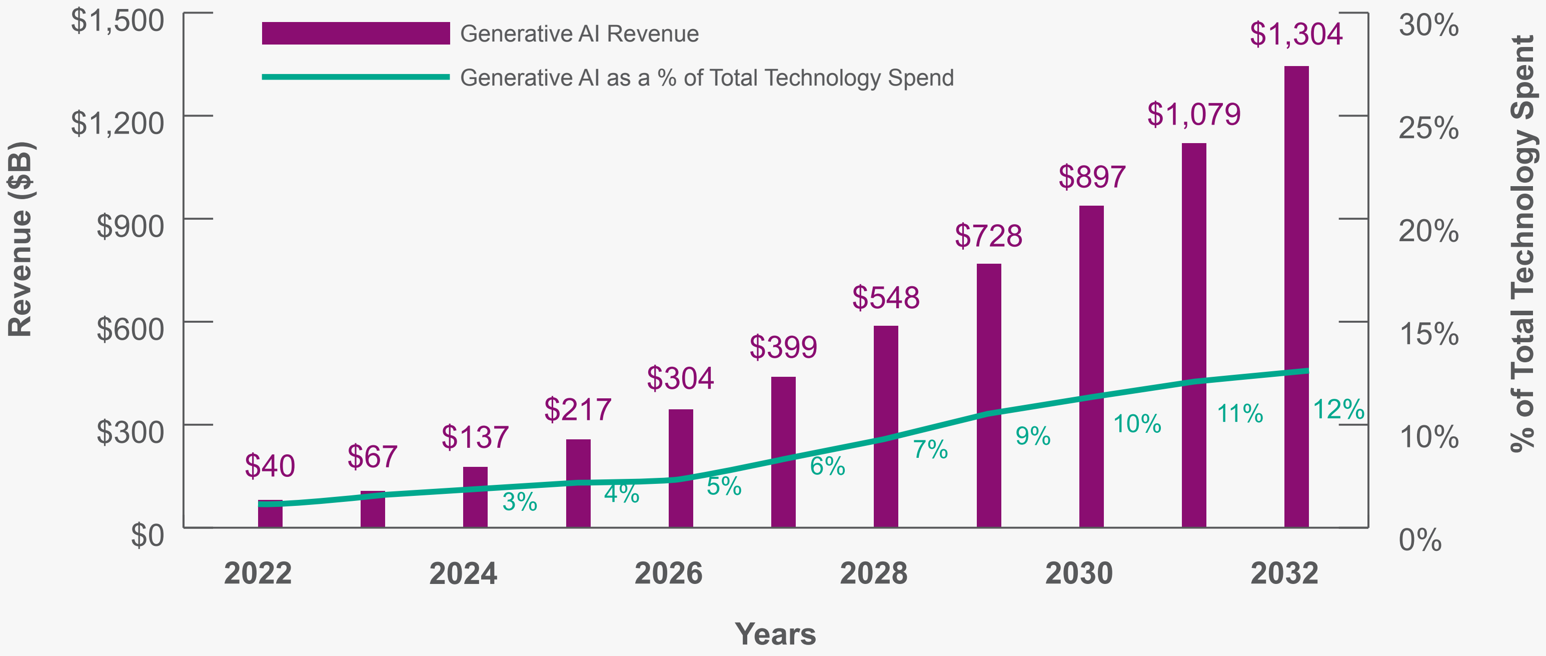
This IMDC Customer Solution Brief provides an overview of the outlook, drivers, opportunities, and infrastructure challenges created by the boom in generative AI.

AI – The Next Generation

Artificial Intelligence [AI] has made extraordinary advances over the last year, driven by the phenomenal uptake and interest in generative AI apps such as ChatGPT, DALL-E, GitHub Copilot and Stable Diffusion, which can create impressive images, answer complex questions, write essays, create websites, and even write computer code, all in answer to a short non-technical text query.

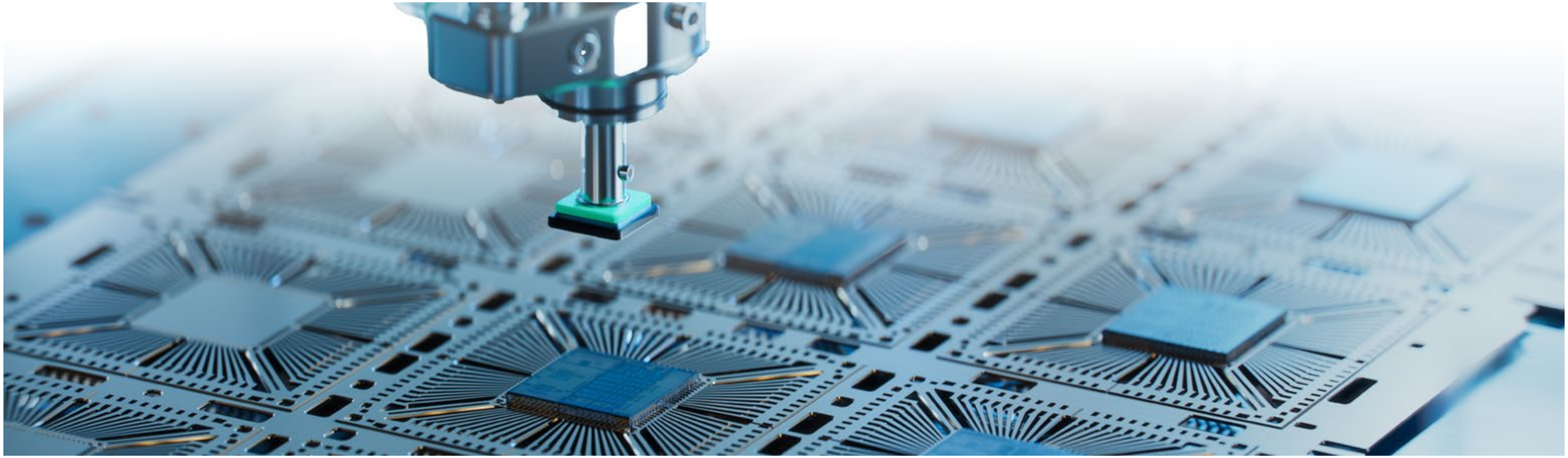
Traditionally, AI systems have been trained on large amounts of data to identify patterns. Generative AI goes a step further, using complex systems and models to generate new outputs in the form of images, text, or audio using natural language prompts. This new model promises massive social and economic value due to its ability to handle multiple repetitive tasks quickly and accurately, combined with this ability to break down communication barriers between humans and machines.

There is inevitably widespread debate over the total economic value and timeframes for the growth of generative AI, but the numbers are huge. Goldman Sachs estimates a value of almost $7 trillion (7% of global GDP) within a decade, while McKinsey estimates that it could add the equivalent of $2.6 trillion to $4.4 trillion annually (3-4% of global Gross Domestic Product). This would increase the total current forecast economic impact of broader AI non-generative applications by between 15 and 40%.



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.

New Chips, New Challenges



As well as accelerating innovation and changing the way people work, the new technology will have a huge impact on IT infrastructure. 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. These training phases then need to be repeated as new data is added.

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. If the model is for general use, this 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 centralised public cloud model that currently supports most applications. 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.

Demand Driver

Capacity and power, as demanded by AI apps, is the biggest issue facing the industry today. Enterprises, AI startups and cloud service providers are racing to secure data center capacity for their AI workloads, with the hyperscale clouds leading the pack. This is happening fast. Investment bank and analyst firm TD Cowen reported "a tsunami of AI demand," with 2.1 GW of data center leases signed in the US (a fifth of current total supply) in Q2 2023.

It is hard to quantify the scale of the challenge. At the largest and most compute-intensive end of the generative AI spectrum, we should be prepared for new and more intensive models and versions. At the smaller scale end of the spectrum, we can still expect at least a doubling in generative AI requirements year on year. As we can see from analyst reports, precise growth levels are hard to predict, but the variations will make a big difference.

- Taking current data center capacity compound growth at a relatively modest 15%, capacity will double in around 5 years and quadruple in 10.
- If generative AI raises data center CAGR to 20%, capacity will need to increase 2.5 times in 5 years and 6 times in 10 years.
- If generative AI boosts data center CAGR to 25%, capacity will need to increase over 3 times in 5 years and 9 times in 10 years.

AI In Action

A Large Language Model (LLM)

ChatGPT 3.5 has 175 billion model parameters and was trained on over 500 billion words of text. According to one article, this recent class of generative AI model requires a ten to a hundred-fold increase in computing power to train models over the previous generation. This doubles overall demand every six months. To train a ChatGPT 3.5 model requires 300-500 MW of power. Currently, a typical data center requires 30-50 MW of power. One of IMDC's larger campuses, in Northern Virginia, has capacity for 10 data centers on it; the whole of the power load for this campus would be required to train ChatGPT 3.5.

A New Healthcare Application

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 since it was first built in 2018 has been steep. It began 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. In less than two years a petaflop of processing power will be needed.

Research AI

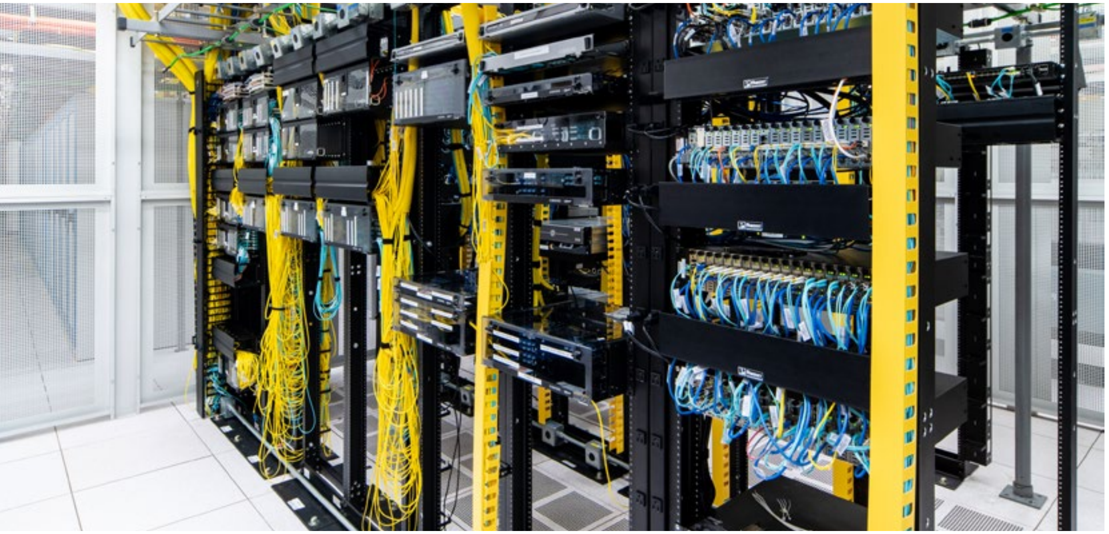
The Computational Research Accelerator department at Arizona State University was running out of network ports, space and power for 'Agave', its supercomputer, so they built a new supercomputer called 'Sol' in 2022 in one of IMDC's Phoenix data centers. Sol is a Dell-built system spanning 178 nodes, it uses AMD Epyc 7713 CPUs (consisting of around 18,000 cores), with the bulk of the nodes carrying 512GB of memory and five large-memory nodes equipped with 2TB. It has 56 GPU nodes with quadruple Nvidia A100 (80GB) GPUs each and 4 nodes with triple Nvidia A30 (24GB) GPUs. The system is networked with Nvidia's 200GB/s HDR InfiniBand and supported by 4 Petabytes of Dell BeeGFS scratch storage. The R&D potential of Sol is extremely exciting, and a steep physical growth curve is anticipated.

Design Impacts

Generative AI training takes place in specialised High-Performance Computing (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 30-50 kW/rack or higher as opposed to the traditional 5 kw/rack.
- Robust backup power for 24/7 operation of high-powered infrastructure - which is driving innovation in low carbon solutions such as battery, hydrogen, and nuclear power to ensure 24/7 uptime for applications.
- Modular design which can cope better with extra complexity and datasets as 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.



- Advanced cooling for the heat-intensive training phases and to extend the lifespan of expensive hardware. While many facilities still offer air cooling, liquid cooling for higher density computing has already begun, as it is significantly more efficient. The industry has been moving there slowly because it was more expensive for lower density installations. However, generative AI, with its much higher-density needs, makes liquid cooling much more cost-effective and conversions will become widespread.
- Faster server innovation. Chips are key to this, and work is already underway to develop new, more efficient, designs. These include, for instance, the Meta Training and Inference Accelerator (MTIA), a new custom chip that Meta designed in-house. Meta claims that the MTIA will be twice as efficient as the GPU chips used today in most current AI infrastructure.

More traditional, but critical, features for generative AI in the data center are:

- Connectivity. High-bandwidth network ports are needed for data upload, exchange and inference. Curated datasets come from all sorts of places, so international multi-Gigabit pipes are a must.
- Low Latency connectivity to inference points. While this is less critical for LLMs performing a linguistic or image-related task, it will be critical for real-time applications and immersive AI such as AR/VR
- Security and certifications. Generative AI is an inventive technology and should only be operated in line with certified protocols to safeguard generative models and proprietary data

Energy & Efficiency

Considering the exponential growth in power requirements to support generative AI, the data center industry will need to look at new generation sources at scale and within a tight timeframe. The industry and its customers can expect a great deal of public scrutiny during this period. Only sustainable and responsible energy sourcing and highly efficient operations and equipment management will ensure that AI applications are widely accepted.

IMDC sites now under hourly clean energy tracking represent over 75% of the company's annual electricity consumption and multiple days of 100% carbon-free operation have been reported at sites with local renewables suppliers.

Climate Commitments

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. The challenge to the industry is clearly not whether they can be achieved, but how. Most hyperscalers and a growing number of colocation providers have been growing the green grid and eliminating carbon. Hyperscalers are already the biggest buyers of renewables in the world, and the more sustainable colocation providers are up there with them. IMDC has been offering its customers 100% renewable power since 2017, and the Iron Mountain group is now one of the top 20 renewable buyers in the world.

Total Decarbonization

IMDC has gone a step further, targeting not just 100% renewables but 24/7 carbon-free power use by 2040. To supplement on-site generation, we are now matching a growing portion of our site electricity use with local clean power generation every hour of every day. This ambitious new approach, pioneered by Google, will in time replace the current year-by-year renewable Power Purchase Agreement model. IMDC is currently the only global data center provider to have committed to this total decarbonization model.

Embodied Impact & E-Waste

More AI-specific facilities and denser high-performance compute architecture will create significant challenges to reduce embodied carbon impacts.

Low-Carbon Builds

AI will undoubtedly accelerate the rate of data center builds. These new facilities, many of which will be customised for AI use, will need to demonstrate embodied carbon reduction. As more building takes place to support the new distributed AI-driven facilities, these need to be models of low-carbon construction. Standards and models for this are already in place, such as the BREEAM low-impact construction standard, which covers everything from community engagement and heat sharing to use of recycled materials and integration of on-site renewables.

Tackling The E-Waste Stream

Sixfold growth in infrastructure means a sixfold growth in equipment, much of it with a faster refresh/replacement rate. Effective management of equipment to optimise performance and reduce waste will be a critical selection factor.

E-waste is one of the fastest-growing waste streams in the world. By 2030, annual e-waste production is on track to reach a staggering 75 million metric tons. Global e-waste is thought to hold roughly $60 billion-worth of raw materials such as gold, palladium, silver, and copper. However, just 17 percent of global e-waste is documented to be collected and properly recycled each year.

The impact of ICT equipment needs to be borne in mind when planning. For both efficiency and impact reduction, AI providers will need to check that IT asset lifecycle optimization and recycling, remarketing and secure disposal are available. Iron Mountain has changed the shape of its business to address this issue. Its Asset Lifecycle Management (ALM) division, which now covers 32 countries, sanitizes over 3 million drives a year and has generated over $1 billion for clients via remarketing and recycling.

Part of The Solution?

AI may also hold part of the answer to this problem. AI algorithms can achieve upwards of 90% accuracy in identifying components for recycling, while eliminating the health risks via robotic sorting. It could be the key to making e-waste recycling faster and more affordable.

IMDC opened the first BREEAM-accredited data center in North America in Phoenix, Arizona, in early 2022, and committed to make all new multi-tenant data centers BREEAM certified by 2025.

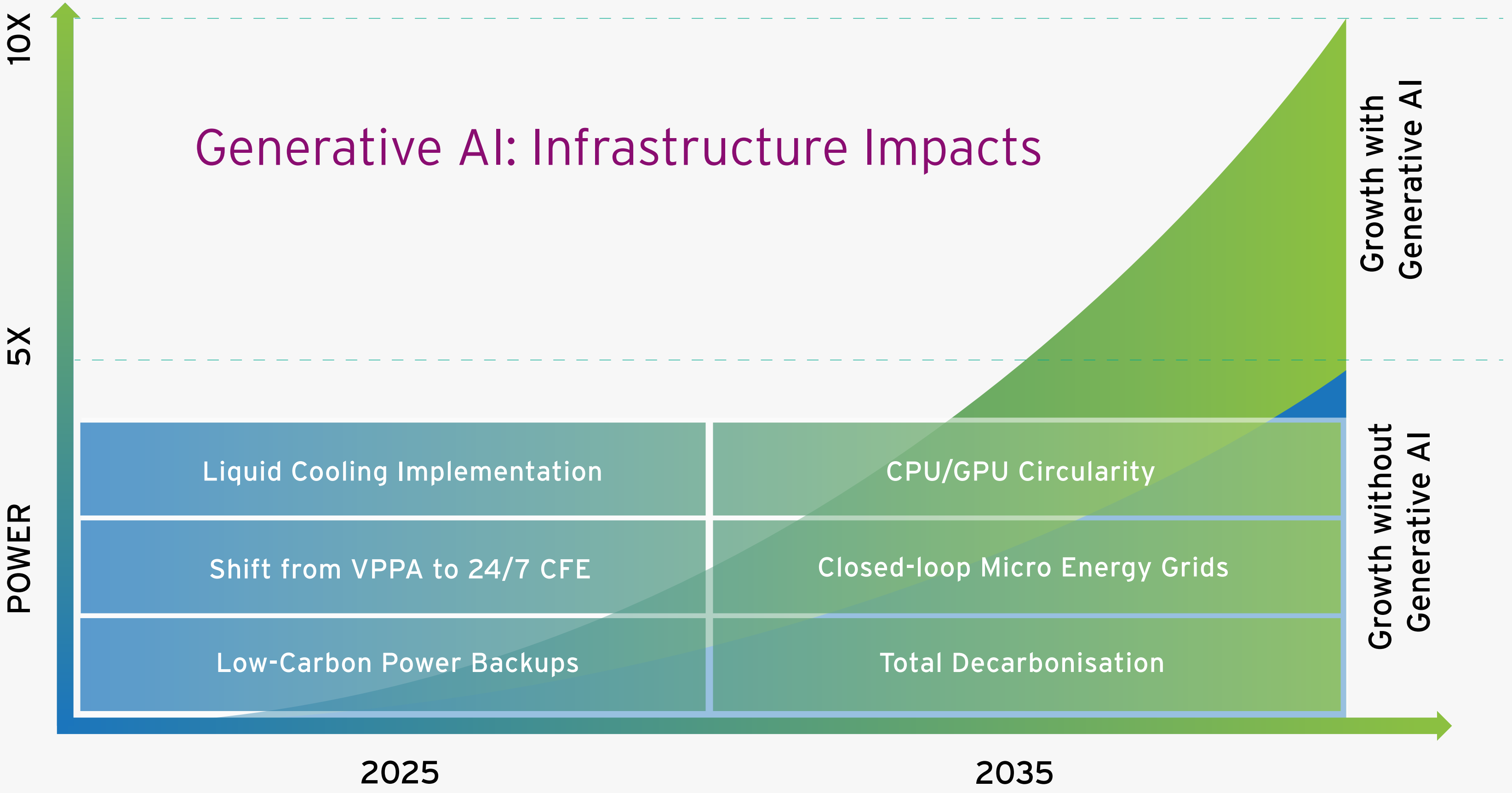
Maximising Opportunity, Minimising Impact

Maximising Opportunity, Minimising Impact

Generative AI is the latest and most celebrated application in a long line of AI-powered applications stretching back to the defeat of Gary Kasparov by Deep Blue in 1997. It promises to deliver immense economic value over the coming decade, but will also consume immense amounts of power.

Many generative AI applications can be hosted in a specialised shared facility. Different models have different infrastructure requirements, but all share the need for high-density power, advanced cooling and modular design.

The scale of the power challenge does not mean it cannot be overcome. In an era in which Big Tech has displaced many oil giants in the list of the world's largest companies, innovation has accelerated and serious fact-based commitments have been made by industry leaders to tackle the climate crisis.



Due to the power-intensive nature of generative AI GPUs, AI-ready data centers are being designed with High-Performance Computing features such as high-power-density, modularity and advanced liquid cooling and low-carbon backup power. In the medium term, the exponential growth in power usage will also drive a race for GPU efficiency, a shift from VPPAs to genuine carbon-free power and a huge new secondary market for CPUs and GPUs.

References

AI - The Next Generation

Goldman Sachs on growth in generative AI value:

https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

McKinsey estimates:

https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#key-insights

New Chips, New Challenges

DCF/ChatGPT power consumption estimate:

https://www.datacenterfrontier.com/sponsored/article/33005561/nautilus-data-center-design-from-legacy-to-modern

T D Cowen on data center demand:

https://cowen.bluematrix.com/sellside/EmailDocViewer?encrypt=78a0ca89-33f7-4a2a-9b79-9bc964666934&mime=pdf&co=cowen&id=world-cowen-morningnotesdistribution@cowen.com&source=mail

Githib on increase in compute power required for new models:

https://github.blog/2023-04-07-what-developers-need-to-know-about-generative-ai/

AI in Action

Arizona State University case:

https://www.hpcwire.com/2022/07/13/asu-launching-sol-supercomputer/

Design Impacts

Modular design efficiencies:

https://www.businesswire.com/news/home/20210225005543/en/Global-Modular-Data-Center-Market-2021-to-2026---Growth-Trends-COVID-19-Impact-and-Forecasts---ResearchAndMarkets.com

Meta Training and Inference Accelerator efficiencies:

https://ai.meta.com/blog/meta-training-inference-accelerator-AI-MTIA/

Energy & Efficiency

Hyperscalers renewable take-up:

Hyperscalers are already the biggest buyers of renewables in the world,

Emissions calculation tools for developers:

Green algorithms

CodeCarbon

Embodied Impact and e-Waste

Statista on e-waste:

https://www.statista.com/statistics/1067081/generation-electronic-waste-globally-forecast/...and collection/recycling: https://www.statista.com/statistics/1066948/share-of-electronic-waste-disposed-globally/

Potential for AI-powered e-Waste sorting:

https://www.sciencedirect.com/science/article/pii/S0956053X20302105

About Iron Mountain Data Centers

Iron Mountain Data Centers operates a global colocation platform that enables customers to build tailored, sustainable, carrier and cloud-neutral data solutions. As a proud part of Iron Mountain Inc., a world leader in the secure management of data and assets trusted by 95% of the Fortune 1000, we are uniquely positioned to protect, connect and activate high-value customer data. We lead the data center industry in highly regulated compliance, environmental sustainability, physical security and business continuity. We collaborate with our 1,300+ customers in order to build and support their long-term digital transformations across our global footprint, which spans three continents.
IRONMOUNTAIN.COM/DATA-CENTERS

To discuss your plans and requirements please contact us on:

US: +1 833 476 2656
NL: +31 800 272 4433
UK: +44 844 417 8379
DE: +49 800 408 0000

Email: datacenters@ironmountain.com

IRON MOUNTAIN®
DATA CENTERS