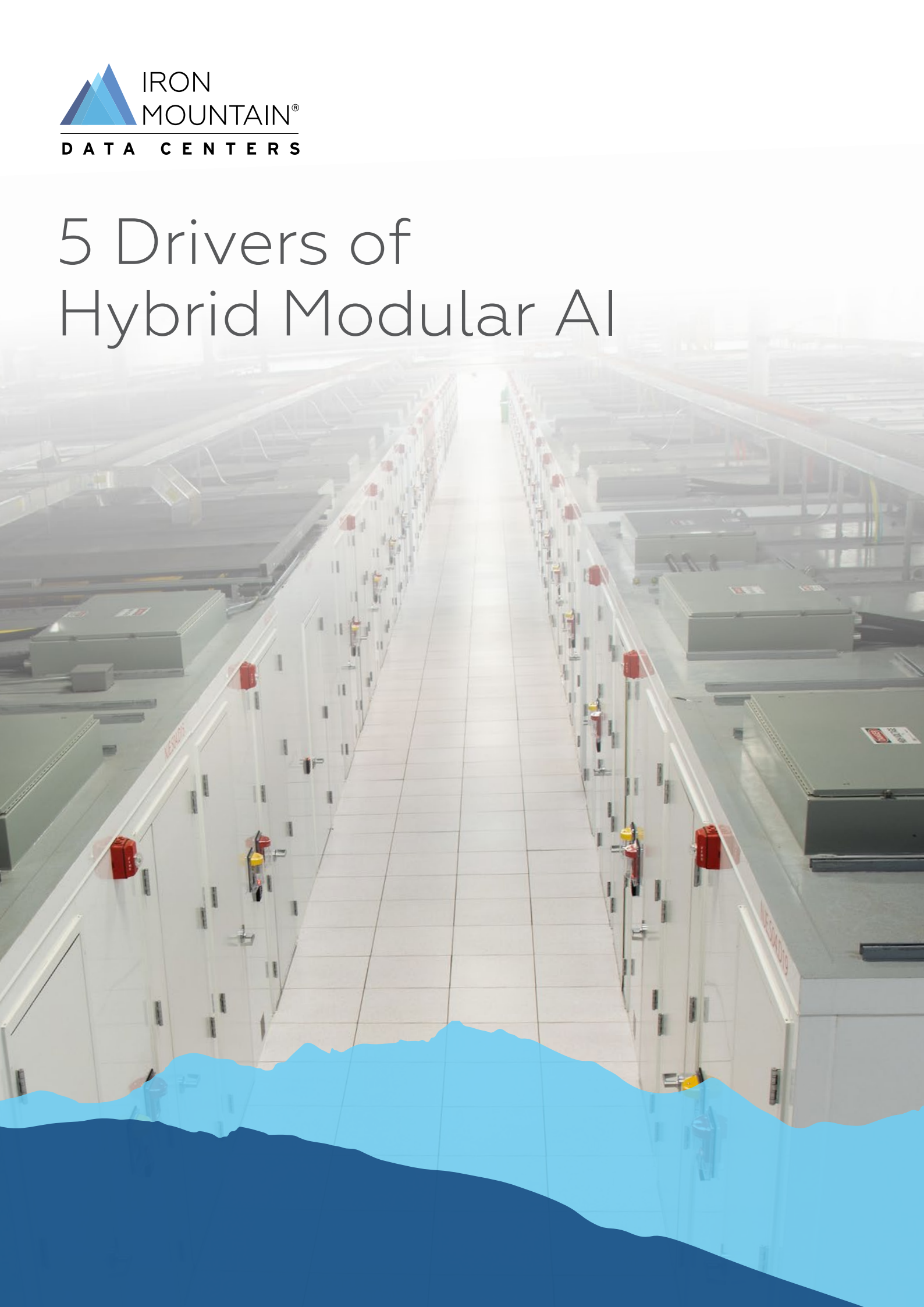
IRON MOUNTAIN®
DATA CENTERS

5 Drivers of Hybrid Modular AI



IRON MOUNTAIN®
DATA CENTERS

5 Drivers of Hybrid Modular AI

While GenAI investment has been dominated by hyperscalers to date, with an estimated spend of over $200 billion in 2024 (according to Goldman Sachs), many organisations are focusing on creating their own AI development environment. Private AI environments address the pressing need for specialist sectoral applications. They are more customizable and secure for sensitive data and sector-specific compliance and evolving regulation. But they are also expensive to run, energy-intensive, and complicated to manage.

These requirements are driving a new surge in interest in the module. Modular AI environments are an increasingly popular solution, particularly for digital businesses that are looking for a hybrid enterprise/AI solution and a smart migration path to broader AI deployment. Aspects of modular design that were sometimes considered over-specified for traditional rack loads are becoming must-have features for AI deployments.

This short e-book sets out the key reasons behind the growing popularity of modular construction for running AI either in isolation or side-by-side with existing non-AI infrastructure:

1. Density
2. Flexibility
3. Efficiency
4. Security
5. Scalability

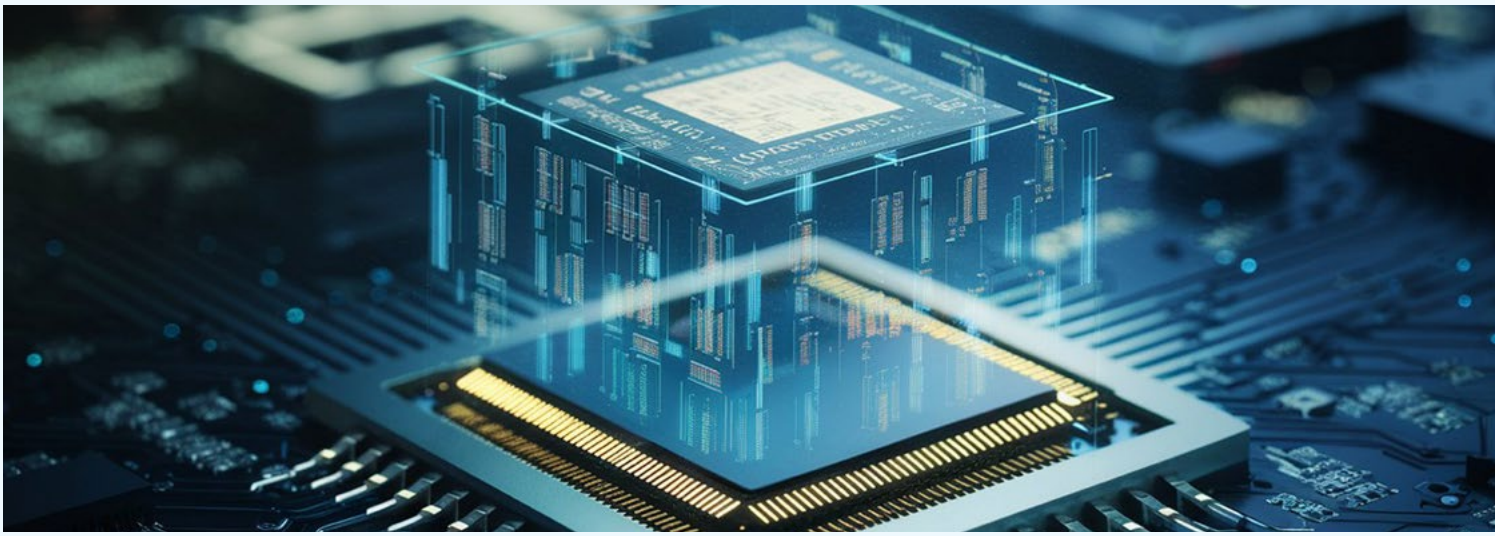
According to IDC, AI could have a cumulative global economic impact of $19.9 trillion by 2030 and drive 3.5% of global GDP by the same year. Hundreds of billions of dollars are being invested in the components and physical infrastructure that will run AI, by public and private cloud providers, governments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

1 Density

There are several reasons for the increasing appeal of modules, and the key one is power. Modules are perfect for more power-intensive deployments that need to be spun up quickly and optimized for specific workloads.

Denser workloads need denser cabling configurations, heavier busducts, cable trays and fiber guides. These are integrated with module design, improving manageability. Freestanding, floor-supported aisle containment systems handle this well while reducing risks during installation. They can also be assembled quickly to create secure aisle containment pathways that can accommodate a wide variety of cabinets and configurations.

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 require 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. This level of cooling complexity is generally very well handled within modular architecture.



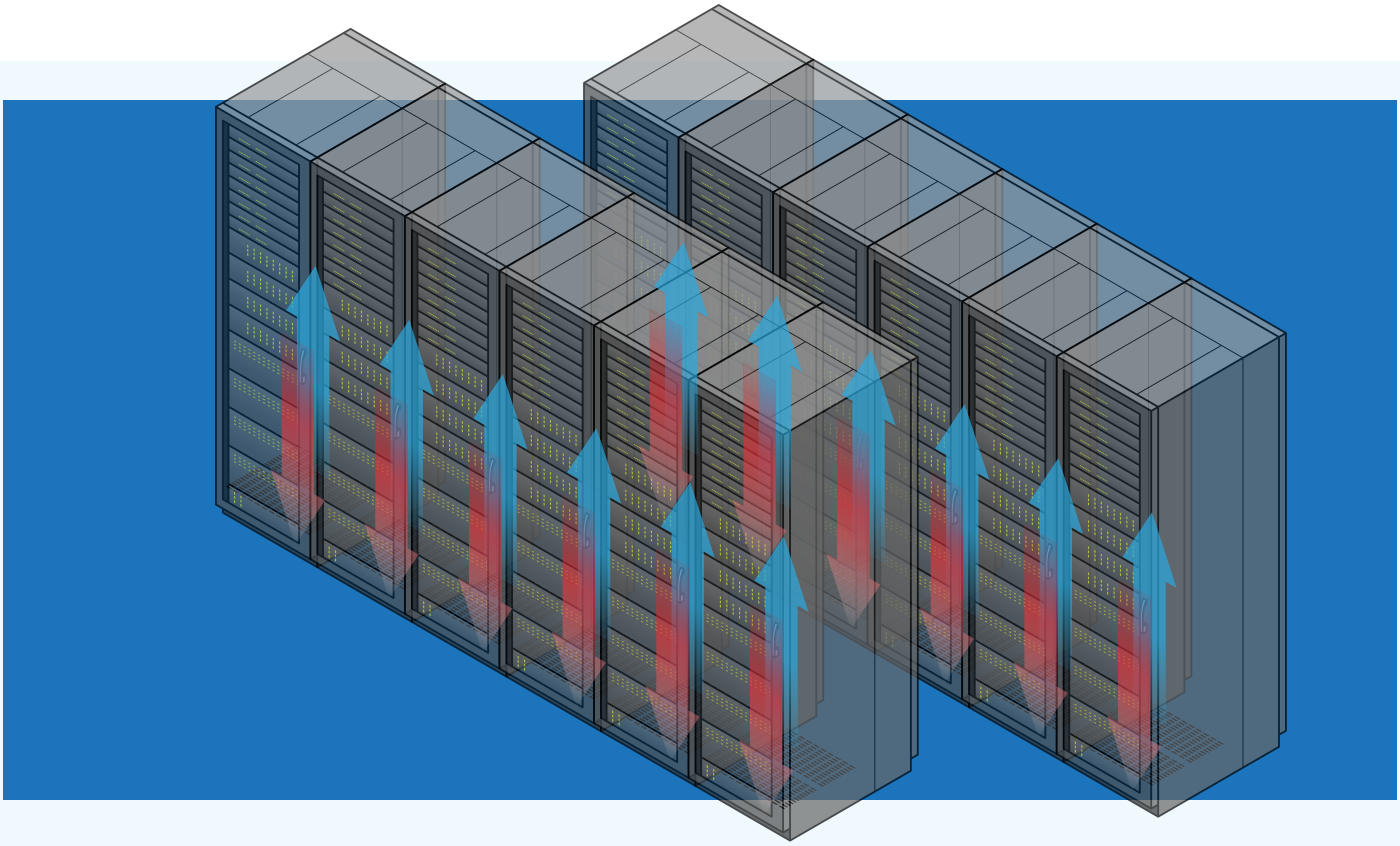
Generative AI models use graphics processing unit (GPU) chips which require 10–15 times the energy of a traditional CPU, because they have more transistors in the arithmetic logic units. This requires more compute, network, and storage infrastructure than any previous technology.

2 Flexibility

Cooling is where the flexibility of module design comes into its own. 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.

This is not a constraint for modules. In a module you can run varieties of AI side by side with less power-intense traditional workloads. Cooling comes right up to the racks under the floor and all you need to do is take out an air header and replace it with chilled water. So if you want to buy in some new AI servers, moving from air-cooled equipment to liquid-cooled, a module can run both, or even do a mix. This is a strategic and planning advantage too, as it makes the modular approach an ideal route for businesses planning incremental AI migration.

As well as cooling versatility, a floor-supported containment architecture provides the flexibility needed to adapt to changing electrical setups. Cabling configurations, fiberguides and busducts can be reconfigured with height-adjustable outriggers or cantilever arms, and you can add material components and supports without disrupting the live environment. Configurations and specifications are changing fast in the current market, and customers need architecture that can keep ahead of that change.



In modular designs, cooling comes right up to the racks under the floor and it is straightforward and quick to remove an air header and replace it with chilled water.

3 Efficiency

Modular deployments are extremely efficient, offering a PUE of 1.2 or less. Designed to optimize space and energy usage, and with a large bank of deployment know-how to back this up, they enable users and operators to make accurate predictions of operational PUE at the commissioning stage.

Of course, efficiency is dependent on background infrastructure and effective monitoring systems, and the sustainability of the overall facility is key. A sustainable operator with optimised energy overheads and ambitious sustainability targets and certifications improves Carbon Usage Effectiveness (CUE) even more. Considering the intensive power demands of AI arrays, every little bit helps. Key considerations for CUE include integrated DCIM software with power usage analytics, renewables access (both on-site and contracted via the local grid), and heat reuse capability.

AI can also play a growing role in improving facility efficiency, with digital twin solutions such as the DCverse AI-Powered Cognitive Digital Twin in the IMDC Singapore facility delivering optimized control policy recommendations for cooling which can improve PUE by several percentage points.

"Iron Mountain gives us greater visibility into our power consumption. We get monitors and tools that allow us to know more about the health of our data center. Those are tools we couldn't have built ourselves."

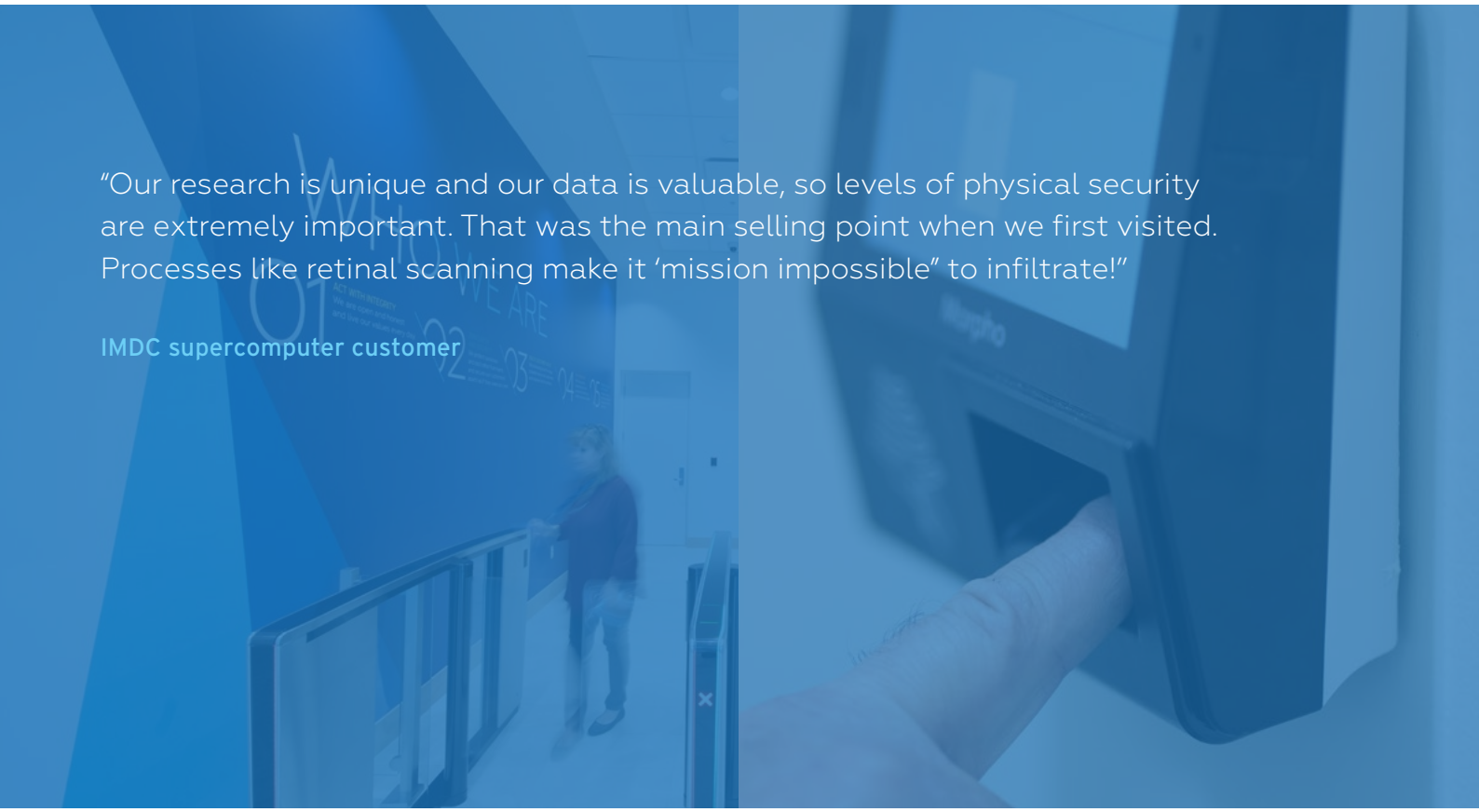
The Computational Research Accelerator department at Arizona State University was running out of network ports, space and power for 'Agave', its supercomputer, so they built a new modular supercomputer called 'Sol' in 2022 in one of IMDC's Phoenix data centers. The R&D potential of Sol is extremely exciting, and it is one of the top-performing supercomputers globally

4 Security

The quality and exclusivity of training data is the key to successful training and competitiveness. For this reason security, privacy and resilience are key concerns. In fact, data security is one of the primary drivers of private AI deployments, giving organisations far greater control of their data in the face of any potential public cloud lock-ins or leaks as well as evolving national regulations.

Modules deliver additional layers of security and privacy. IMDC customers that run AI systems using modular architecture include medical, defence, crypto and academic R&D. They all require the highest levels of data and equipment security and resilience for their proprietary data, and using modules gives them extra layers of security such as cards, biometrics and cameras.

A module is effectively a complete data center in a box. It is physically separate from neighboring equipment and offers the complete privacy of a suite. The fully redundant architecture of modules also supports stringent compliance certifications such as NIST 800-53.



"Our research is unique and our data is valuable, so levels of physical security are extremely important. That was the main selling point when we first visited. Processes like retinal scanning make it 'mission impossible" to infiltrate!"

IMDC supercomputer customer

5 Scalability

Perhaps the most attractive feature of modular infrastructure is its easy scalability. One of the most predictable features of any AI array is that it will grow exponentially. All of IMDC's AI and HPC customers at least double their requirements year-on-year, as millions of datasets become billions, and teraflops become petaflops.

This is where modules really stack up, as new compute resources can be deployed quickly and smoothly across multiple paired walk-in on-site modules. In a supply-constrained and highly competitive market, the advantage of rapid, resilient scalability is hard to overestimate.

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 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.

The Road to AI

It is still hard to see exactly what the next generation of killer AI applications will be, but there are huge opportunities for innovators of all shapes, sizes and sectors to develop winning AI-driven solutions based on smaller AI clusters. 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 & cooling

Can your provider offer sufficient power in the right high-density configurations? 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.

Flexibility & scalability

Can you run your AI in tandem with your other needs? And will your deployment be able to scale up fast; your processing power and storage requirements are likely to rise exponentially

Efficiency & 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.

In the right data center, a modular approach ticks these boxes, which accounts for the sharp rise in interest in modular solutions.

If you are interested in finding out more about planning, locating and activating either hybrid or dedicated modular AI infrastructure, please get in touch:

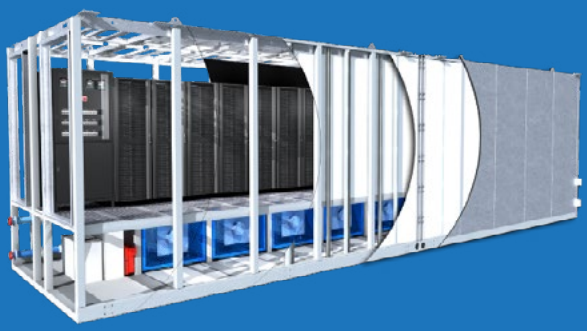
NL: +31 800 272 4433

UK: +44 844 417 8379

DE: +49 800 408 0000

US: 833.IRM.colo

Email: datacenters@ironmountain.com



833.IRM.COLO | ironmountain.com/data-centers

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 90% 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 within our 5M+ SF global footprint spanning 3 continents. For more information, visit www.ironmountain.com/data-centers.

IRON MOUNTAIN®
DATA CENTERS