



AI inferencing infrastructure

4 things to look for

AI inferencing infrastructure

4 things to look for

Larger, clustered, more densely configured data center campuses will be the key to successful delivery of AI applications, but the AI landscape is changing fast.

This Iron Mountain Data Centers (IMDC) ebook gives a quick overview of evolving customer needs, and sets out the key things to look for in a data center partner as you build out accelerated inferencing infrastructure to deliver new AI services.

Contents

3 / Customer needs

Clouds on the horizon
From training to production

6 / What to look for

- ① A global power pipeline
- ② Advanced power & cooling
- ③ Sustainable standards
- ④ Flexibility & focus



Customer needs



Clouds on the horizon

The scale of demand for AI infrastructure is not in question, but the nature of that demand is worth reviewing, as it is entering a transition phase. There are two key factors in this: the growth of neoclouds, and the shift from training to production.



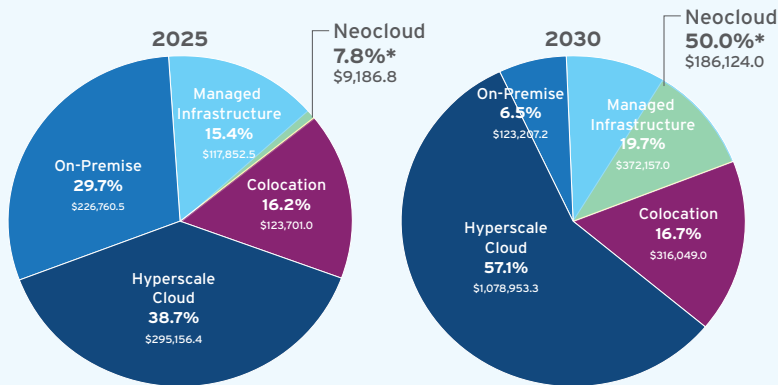
Who is buying AI data centers, and what do they want?

Hyperscale clouds are far and away the biggest customers, and they are consistently increasing their infrastructure investment. But there are others. In addition to growing enterprise and cloud demand, the market is being accelerated by neoclouds, a new generation of specialized cloud computing providers looking for purpose-built high-density facilities.

Cloud-driven growth

Outsourcing IT to accelerated cloud and AI service providers will drive the market, with strong growth in the managed infrastructure segment driven by the neoclouds. By 2030 hyperscale clouds are forecast to grow their share from 38% to 57%, passing the trillion dollar milestone, and neoclouds are forecast to account for almost 10% of the market. As hyperscalers and neoclouds grow this market share, on-premise data centers are forecast to become a far less popular option, with their share shrinking by 80% from 29% in 2025 to just 6% in 2030.

Forecast hyperscale & neocloud demand to 2030



* Percent of total Managed Infrastructure market

Source: Structure Research

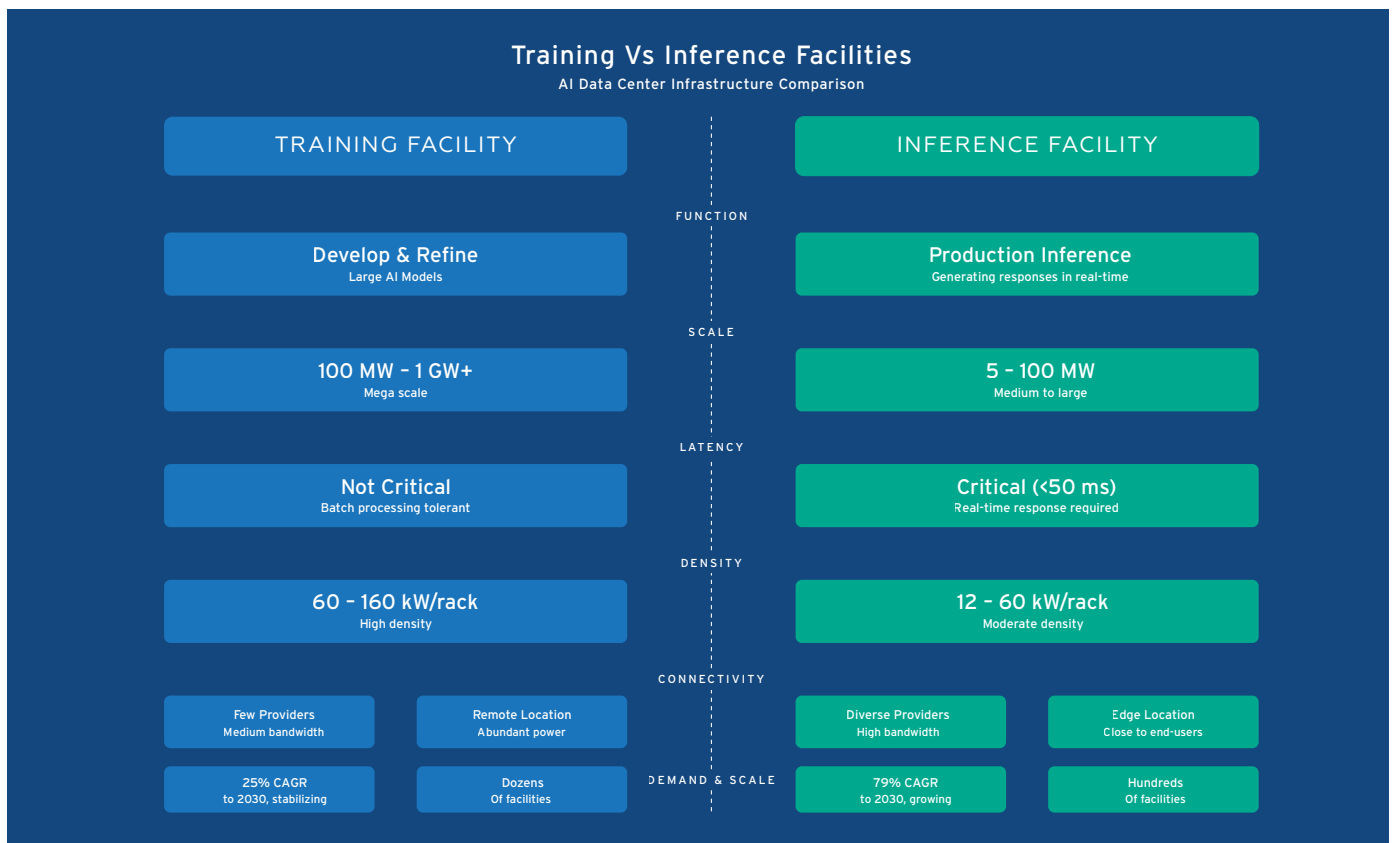
From training to production

Cloud customers are looking for a different type of data center. The highest-profile AI facilities today are vast AI training factories, but the sort of infrastructure that AI needs actually comes in all shapes and sizes. An era of production, aka inference, is now underway, and this has new infrastructure requirements.



What sort of data centers do customers need?

Inference is a recurring need which builds, rather than a one-off training run. Every query and agentic AI application will require inference infrastructure, and it will build incrementally as AI matures and new applications are developed. According to Structure Research, inference infrastructure is forecast to grow at 79% CAGR to 2030, as compared to 25% CAGR for training. This means there will be more inference infrastructure than training infrastructure by the end of this year. By 2030, 80% of all AI infrastructure will be used for inference. There are several key differences between training and inference infrastructure:



Outsourcing the builds

Clouds tend to build and operate their own training facilities, but due to the scale of inference growth they are increasingly outsourcing the development of this new phase of infrastructure to specialist data center operators.

What should they look for in an AI data center partner?

Starting on the next page, you can find four key considerations

What to look for



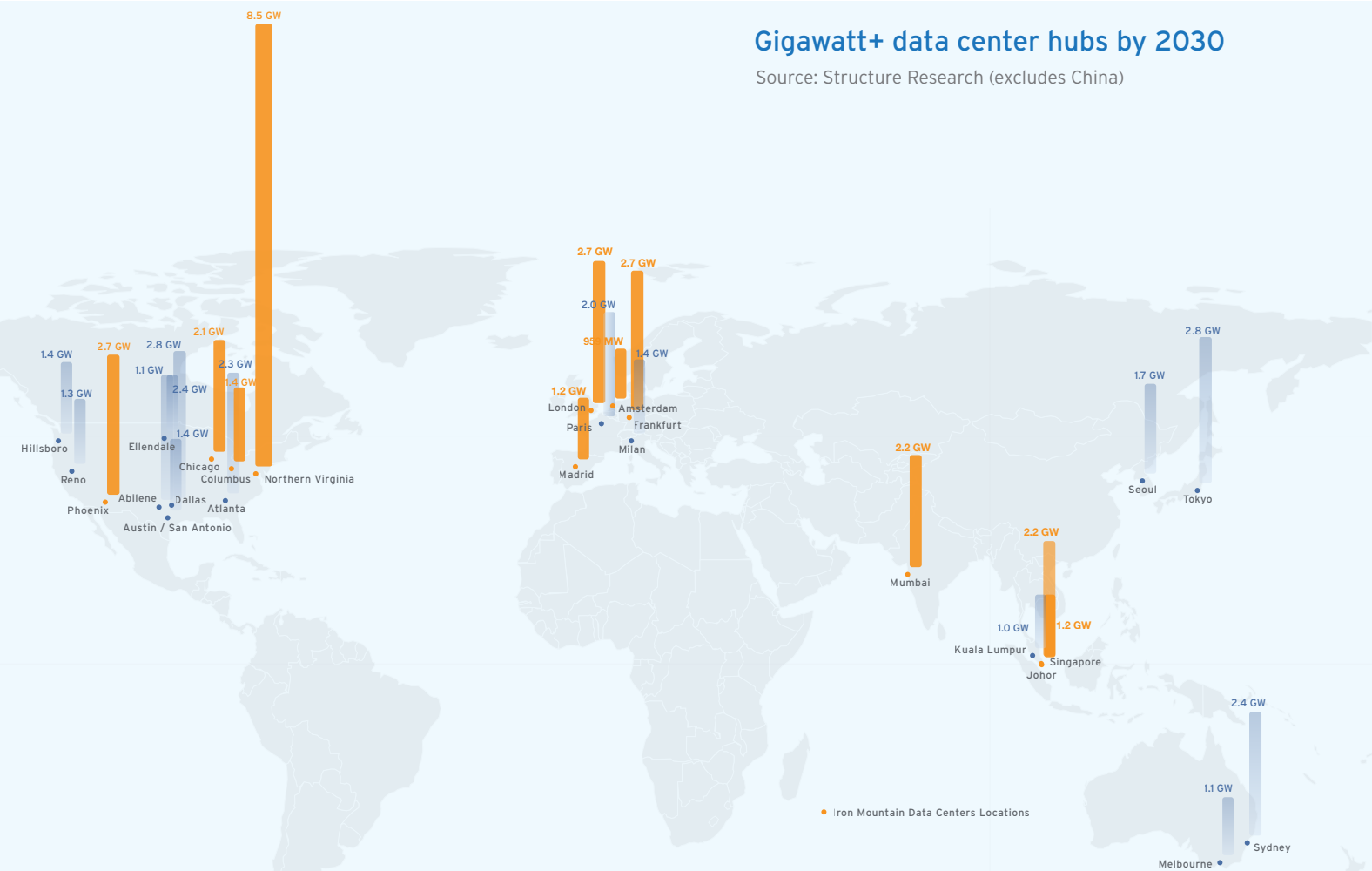
① A global power pipeline

What to look for

Training can be conducted remotely then delivered. Inference cannot. Businesses need local infrastructure to share their data, and that requires a distributed global footprint. Land used to be the primary consideration when siting data centers. Now it is power. As power requirements shoot up and grids in mature hubs become constrained, finding capacity is the overwhelming preoccupation of data center site selection teams.

Gigawatt+ data center hubs by 2030

Source: Structure Research (excludes China)



Total data center capacity in the 25 metros marked here is forecast to grow by 162% in five years, from 21.53 GW (2025) to 56.45 GW (2030) in a mix of Tier 1 and fast-growing Tier 2 “overspill” hubs. This will create 19 new Gigawatt+ data hubs.

Partner profile

Operators with multi-market objectives should look for data center operators with a multi-region power and land pipeline that matches their ambitions. IMDC, for instance, has committed to doubling its development pipeline from our current level of 1.3 GW to 2.6 GW by 2030.

Current IMDC developments include 350 MW of capacity in Manassas and Richmond; 79 MW of capacity in Madrid, one of Europe’s fastest-growing hubs; and 142 MW spread across the fastest-growing hubs in India. With more in the pipeline.

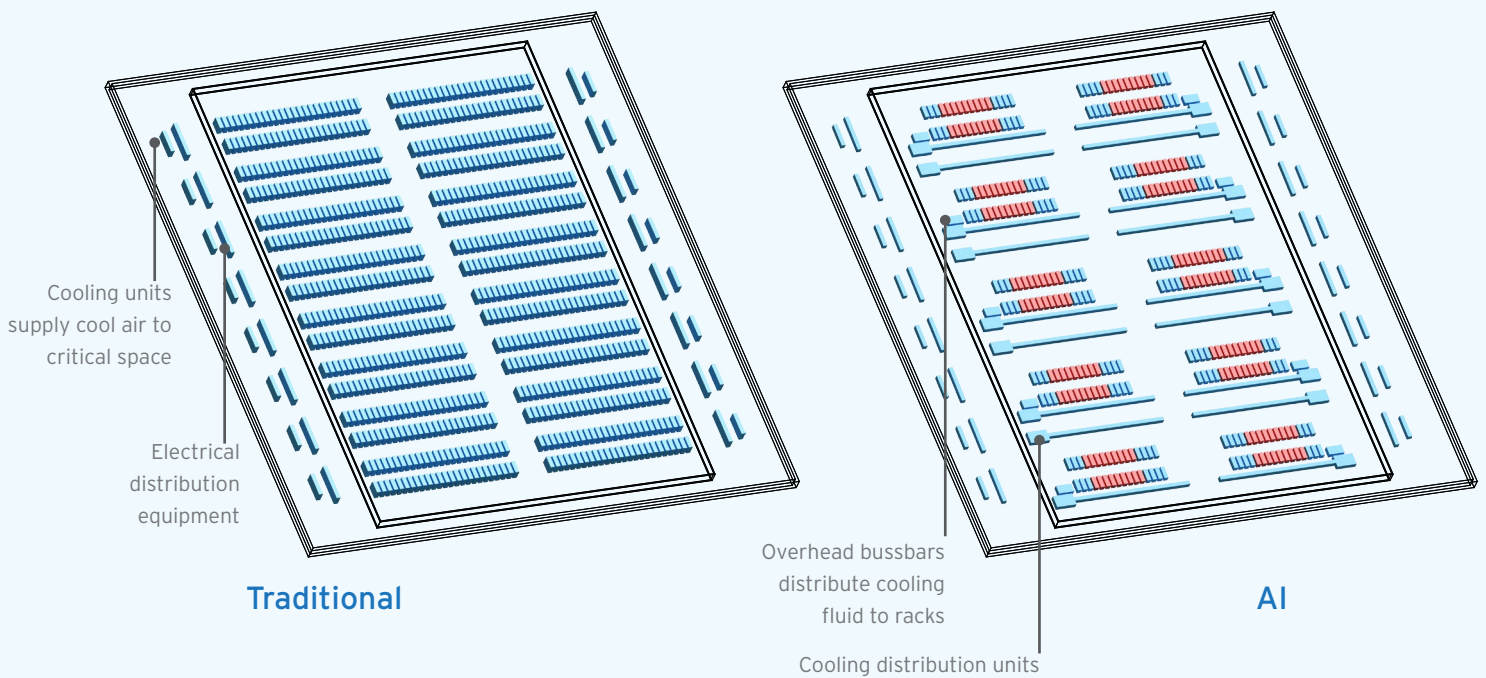
② Advanced power & cooling

GPUs require around 10 times more power density than CPUs, and rising. According to epoch.AI, while compute power is doubling roughly every 5 months, power consumption is doubling every 1.4 years. This power intensity has a rolling impact on design, the main one being the incorporation of liquid cooling, which can pull 3000 times more heat out than air.

AI has transformed data center design

Traditional data centres are packed with network racks. These are typically air-cooled and need 5KW to 10KW of electrical power

AI racks are liquid cooled, and require more than 10 times as much power



A flexible air-cooled traditional data hall would support 12 MW, whereas the same size of hall can support 32 MW with liquid cooling. This ratio will change once more with new running temperatures, power and cooling solutions.

Active Rear Door Heat Exchangers (50kW-70kW) are becoming much more widespread for medium-intensity AI workloads including the H100 GPUs. A step up from this is Direct-to-Chip Liquid Cooling, which can support power densities of between 50kW and 200kW.

Power tools

Power needs to be managed differently in GPU data centers. Components need to be larger and nearer to the compute. Redundancy may not need to be 2N, impacting UPS design. And Behind the Meter (BtM) capability is a useful tool when self-generation or load shifting are needed due to grid constraints.

Partner profile

Look for a data center partner with an experienced and well-resourced energy, design and construction team. Reference designs that can deal with high densities through a mix of advanced air and liquid cooling are key.

IMDC runs two standards globally, but often develops tailor-made designs for larger customers. All are validated against energy-efficiency and operational performance criteria. Most of IMDC's AI-ready facilities are currently running at medium-level densities of 12-50 kW/rack, which are suitable for inference. Densities are continually increasing, and 200 kW racks are now being piloted in IMDC Virginia.

IMDC also has experience in BtM design, with Battery Energy Storage System (BESS) projects in Virginia and New Jersey.

3 Sustainable standards

Larger, higher-density infrastructure needs to mitigate its climate impact. Cloud Service Providers are world leaders in the deployment of energy efficiency measures and renewable power. But the speed of the energy transition, grid limitations, and the overwhelming need for more higher densities threatens to undermine ambitious emissions targets.

As the world electrifies and eliminates carbon, a power crunch is taking place. Clear standards, accurate and transparent reporting of embodied and operational emissions, and year-on-year improvements to achieve emission reduction targets are the answer.

New GHG Protocol

Emissions standards are changing and remote offsetting is coming to an end. The new [Greenhouse Gas Protocol](#) stipulates that all clean energy claimed should be local and simultaneous with use. This may cause challenges for some operators in the short term, but it will be a key lever for reducing the real carbon footprint of accelerated data center infrastructure.

Partner profile

Partners will need to track carbon emissions locally and hour-by-hour. In addition to its 100% renewable VPPAs, IMDC's 2022 commitment to 100% clean power tracked

hour by hour by 2040 is unique among global data center operators. 75% of IMDC customer power is now tracked hourly and energy procurement lowers carbon levels via local and on-site renewable sources.

Power and water efficiency are also key. IMDC facility design aligns with the [Climate-Neutral Data Center Pact](#) targets for PUE and WUE.

Builds need to be demonstrably sustainable. BREEAM is a global construction standard which assesses every aspect of the data center build from planning, consultation and communication with the local community to land water and materials use, energy, ecology, innovation, annual operational metrics, transport and recycling. IMDC has adopted the build standard worldwide.



In 2022 IMDC's Phoenix AZP-2 data center was the first data center in North America to earn the BREEAM (Building Research Establishment's Environmental Assessment Method) sustainable design certification. The BREEAM standard is now applied to new IMDC facilities worldwide, with 13 sites either certified or in process.

4 Flexibility & focus

Accelerated computing is just that - an evolution and speeding up of an existing technology. Cloud infrastructure has developed hugely over the last 15 years, and now it is changing faster than ever. New reference designs and power solutions for AI facilities need to take account of these shifts. Operators need to be adaptable for all eventualities. In other words, flexibility and responsiveness are key.

Specification changes, redirects and workarounds are inevitable in such a pressurized technology-driven market phase. Data center operators need to provide reassurance that these and future challenges will be solved quickly, creatively, and sustainably.

Supply chains need to be flexible, but strong. At the permit level, mature but pragmatic relationships are needed with regulators, planners and power providers.

Customer communication needs to be clear every step of the way to delivery.

Look for **aggressive and creative site selection** in multiple regions and sustainability leadership. Also, make sure that the design and construction team not only think outside the box, but redesign the box to suit your project needs. Sometimes designs need to be one-off for particular markets. This can mean provisioning for multiple design and build options and changing plans at the last minute.

A **standards-based approach** is the critical departure point. The best way to provision for different layouts - different temperatures, generator quantities, extra measures for spikes in processing power, data striping, varying capacities and densities - is to overlay them on tried and tested blueprints and calculate the different construction and operational impacts.

Effective on-time delivery is the ultimate measure of success. This requires razor-sharp focus.

Specialization across all key functions enables quicker decision and global scale. These include energy procurement and partnership, site selection, design, construction, sourcing, mechanical and electrical project management, and solution engineering.



To manage large-scale builds which require complex provisioning, customer satisfaction is more important than ever. IMDC uses the Net Promoter Score (NPS) to measure this. As builds have grown in size the company NPS has improved year-over-year, increasing from 54 in 2024 to 56.9 in 2025, with scores of over 50 in every operating region. Anyone who works with the NPS ratings will recognise that these are exceptional levels.

Iron Mountain Data Centers

Protect, connect and **activate** your data

Your data is your advantage. Yet too often it remains untapped: disconnected from systems, underutilized, untrained, and exposed to risk. At Iron Mountain, we help the world's most complex organizations unlock what's possible with data centers that protect, connect, and activate your data like never before.

How? By combining our legacy of trusted security with advanced, cloud-agnostic data centers powered by 100% clean energy - delivering scalable, compliant, sustainable and future-ready solutions. Our global 1.3 GW portfolio empowers customers including leading cloud and AI providers to scale their data footprint, enhance security and unlock greater value.

What can we unlock together?



www.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

