IRON MOUNTAIN®
DATA CENTERS

The future geography of AI infrastructure

Spheres of inference

The future geography of AI infrastructure

Spheres of inference

While the tech refresh rate for chips appears to be speeding up, there are some parts of the accelerated computing revolution that are easier to predict. Because building takes time, data center infrastructure is one of them.

This Iron Mountain Data Centers (IMDC) ebook gives a quick overview of the factors that are driving and shaping the expanding infrastructure footprint that will support high-performance hyperscale and neocloud services around the world.

Contents

4 / The new drivers

The adolescence of AI
Power, production & people

7 / Spheres of infernce

North America Top 10
Europe Top 10
Asia Pacific Top 10

IRON MOUNTAIN®
DATA CENTERS

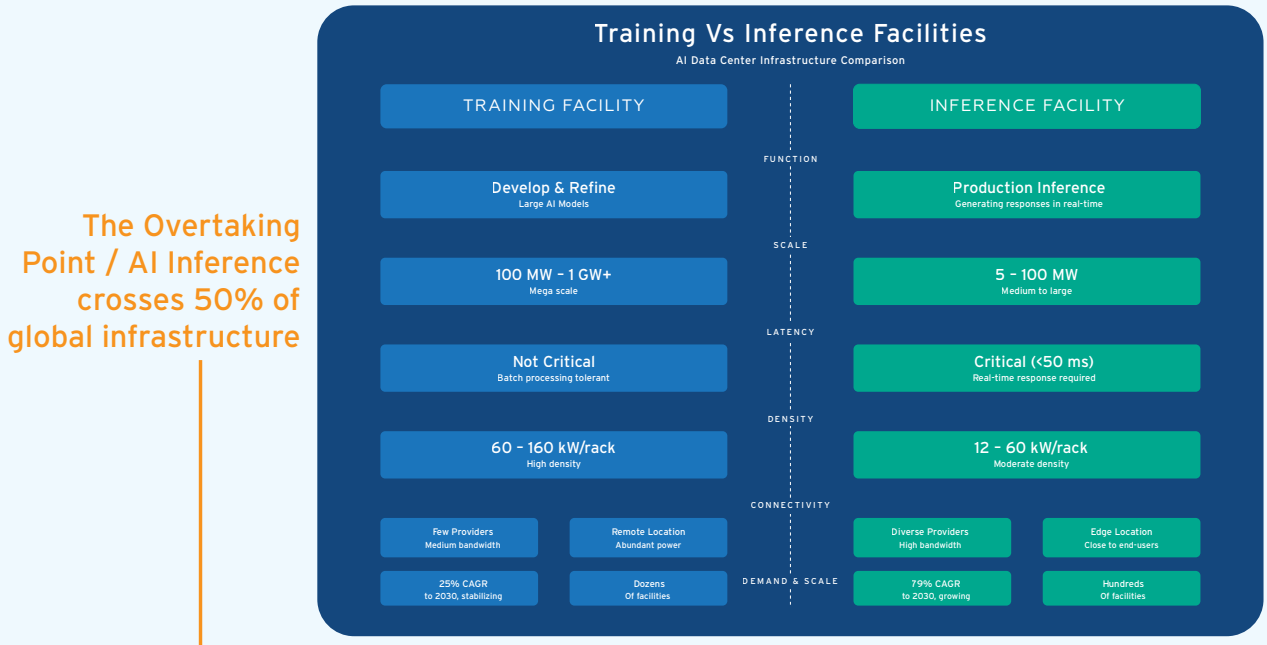
The new drivers

The adolescence of AI

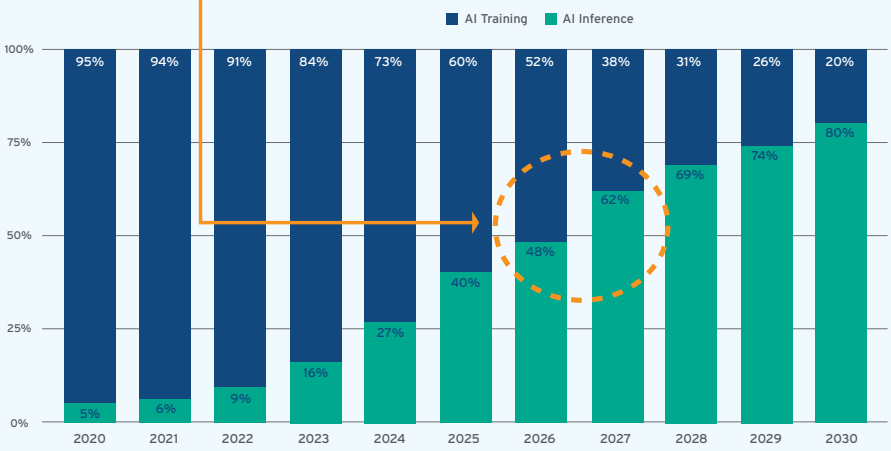
While a lot of complex AI models are still being developed, the first generation of LLMs and agents are now out on the street and earning, having graduated through training to production, aka inference. By its nature, inference will require much more capacity than training, as it will be round-the-clock and cumulative. To support this coming of age, the next few years will see a wholesale shift in the investment in high-density liquid-cooled large-scale infrastructure, away from (but maintaining) training, towards less "mega" more distributed end-user-focused inference hubs.

What´s next for AI?

- The scale of the challenge: According to McKinsey "to avoid a deficit, at least twice the data center capacity built since 2000 would have to be built in less than a quarter of the time."
- Growing AI dominance: About 20% of total data center capacity is already being used for AI, and Cushman & Wakefield estimates that AI will drive $75 billion in data center demand by 2028, pushing this share up to 35% of the total market. By 2030 this figure is likely to be 50%.



The Overtaking Point / AI Inference crosses 50% of global infrastructure



AI TRAINING CAGR 25.4%

AI INFERENCE CAGR 79%

By next year, AI inference infrastructure will account for 62% of all AI infrastructure and continue on a much steeper curve than training, with forecast growth of 79% CAGR to 2030, as compared to 25% CAGR for training. This has huge implications for the extent and geographical spread of AI infrastructure.

Power, production & people

Power availability is the primary dictator, and limiter, of AI´s growth, with particular pinch points in and around densely populated mature markets, struggling with ageing infrastructure and the energy transition. This puts these areas, the traditional locations for clouds and colocation, at a premium.

What are the key drivers for AI infrastructure?

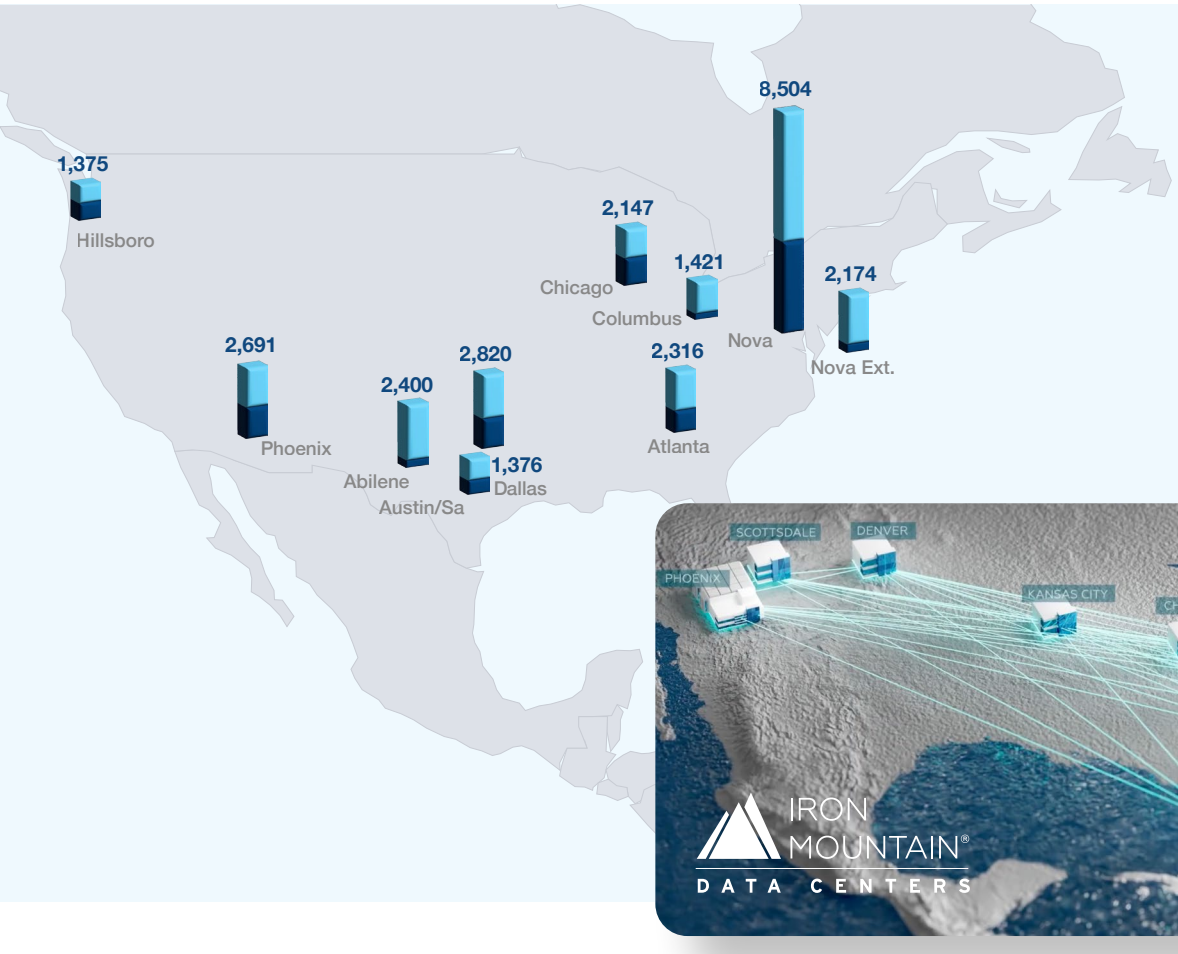
- To support enterprise AI growth, which is forecast to run at 84.9% over the next five years (Structure Research), AI facilities need to be integrated into existing cloud regional architecture, with latency-sensitive inference workloads embedded next to existing cloud on-ramps.
- Cloud zones are getting significantly bigger, and a new regional architecture is emerging. This includes extensions to Availability Zones with much larger capacities. These are optimized for both inference workloads, which are very latency sensitive, and satellite inference + training workloads which are not so latency sensitive.
- Where the strain on existing generation and transmission is too high, fast-growing "spillover" hubs are being formed, where power can be provided in reasonable proximity to existing cloud availability zones.
- AI data sovereignty is now a major geopolitical issue, making training as well as inference investment more geographically diverse.

IRON MOUNTAIN®
DATA CENTERS

Spheres of inference

North America top 10

Where will the North American top 10 AI hubs be in 2030?



- In the USA, Northern Virginia (NoVA) will continue to stand head and shoulders above the rest, reaching 8.5 GW of capacity by 2030, with a sizeable spillover zone referred to in the graphic below as the NoVA Extension (2.2 GW by 2030)
- Other established hubs such as Dallas (2.8 GW), Phoenix (2.7 GW) and Chicago (2.1 GW) will also see massive absorption
- On the Training side, remote GPU "supercluster" locations are emerging in regions offering advantages such as abundant, low-cost renewable energy and cooler climates, like Abilene, TX (1.2 GW), and Ellendale, ND (545 MW)
- The rise of ex-crypto GPU data center providers further contributes to this, with significant capacity forecast in Texas (about 1.9 GW) and North Dakota.

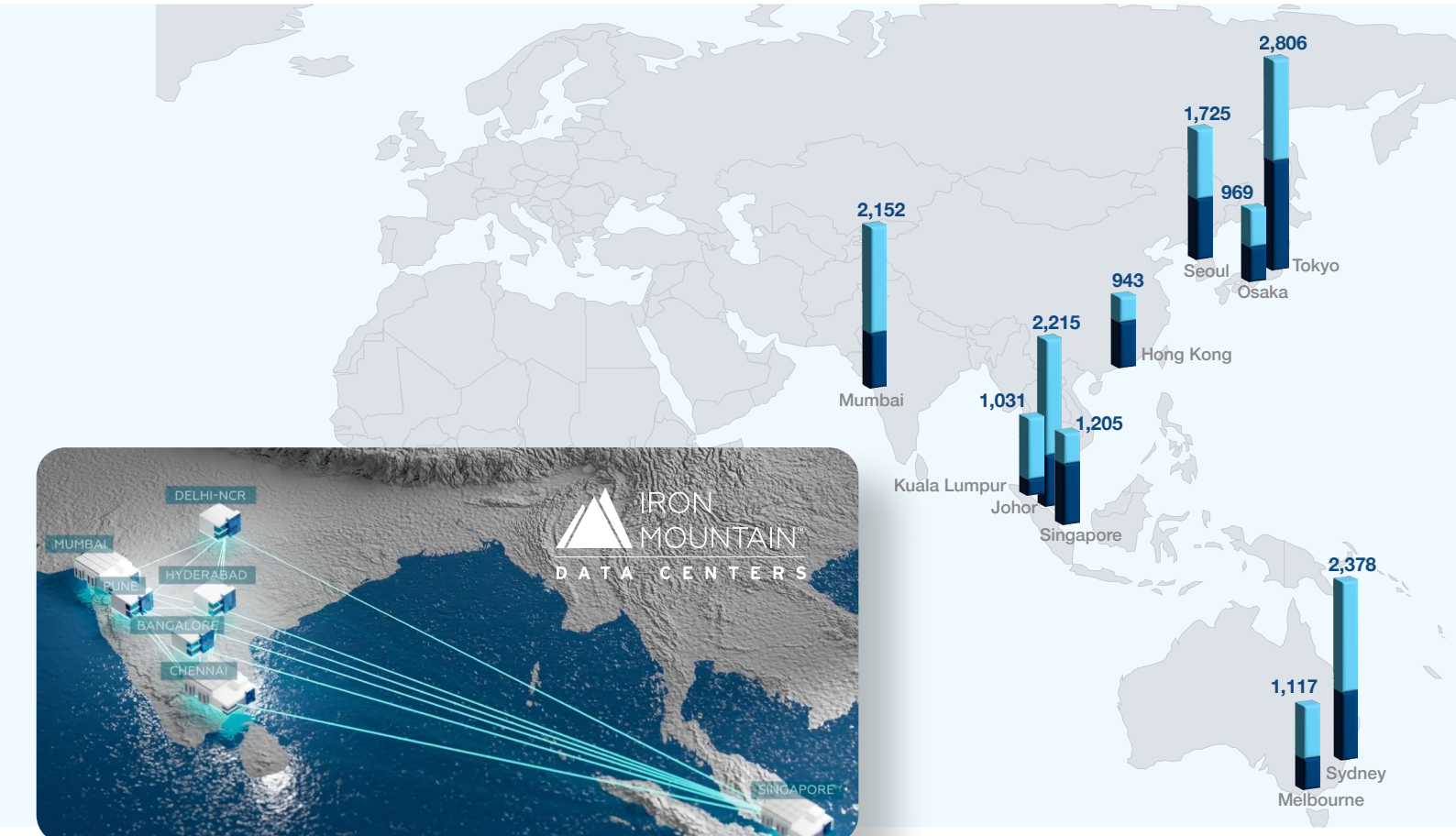
Europe top 10

Where will the European top 10 AI hubs be in 2030?

- Infrastructure growth in Europe is forecast to grow even faster than North America through 2030 (24.7% CAGR versus 20.9%)
- Core hubs will continue to be London (2.7 GW), Frankfurt (2.68 GW), and Paris (2 GW)
- Many new hubs are growing fast in Europe in particular, driven by power constraints, renewables availability, and data sovereignty regulations
- Leading new European AI hubs will include Milan (Italy), Madrid (Spain), Berlin (Germany) and Lisbon (Portugal). These sites offer high power availability, low power costs, renewable energy and proximity to connectivity aggregation points.

Asia Pacific top 10

Where will the Asia Pacific top 10 AI hubs be in 2030?



- Even excluding data from China, where reliable figures are not available, the Asia Pacific region covers a massive area and a huge number of extremely diverse markets
- The current leading markets Tokyo and Sydney will maintain their position in 2030 with 2.8 GW and 2.4 GW respectively
- Seoul (1.7 GW by 2030) and Singapore (1.2 GW by 2030) will continue to grow fast, but not as fast as Johor (2.2 GW) and Mumbai (2.1 GW) which will overtake them
- Johor provides a high-potential overspill market from strategically located but space-and-power-constrained Singapore, and Mumbai is the first of a cluster of high-growth Indian hubs.

Spheres of inference: the future geography of AI

The nature of AI and advanced cloud infrastructure requirements is already driving a major geographical shift, as the industry pivots from focusing on large-scale training, which requires remote, power-abundant and politically significant mega-clusters, toward the widespread deployment of latency-critical inference capabilities, or production hubs.

This pivot requires a fundamentally more distributed architecture, consisting of medium-to-large facilities located close to end-users and integrated into existing cloud availability zones and on-ramps to ensure critical low latency.

The other major geographical shift over the coming years is overspill: While Tier 1 hubs will continue to grow where possible, new dynamic hubs are emerging fast.

From Virginia, to Southern Europe, to Singapore, space and power constraints are accelerating the growth of these "spillover" markets. If they are also renewable-rich, populous, and networked, they will grow even faster.

This growing focus on inference represents a major opportunity for hyperscale-focused colocation and build-to-suit providers like IMDC. Inference data centers require build speed, densely populated locations, diverse bandwidth, and denser interconnection, all of which play to our core strengths. To support and enable AI, we are constantly investigating suitable new locations and expanding our overall capacity pipeline worldwide.

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?

IRON MOUNTAIN®
DATA CENTERS

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