B · METAL

RESEARCH REPORT

A BMETAL RESEARCH REPORT

The Urbanization of AI Compute

Why AI inference is moving back into the city — and what it means for the physical layer of AI.

GLOBAL · NORTH AMERICA · W. EUROPE · LATAM

OUTLOOK TO 2035

PUBLISHED JUNE 2026

CONTENTS

Inside this report

Figures are drawn from public industry research and cited inline; full sources are listed at the end. Cross-references and citations are clickable.

Executive Summary

The AI buildout is splitting in two, and the second half is moving into the city.

Training — the work of building frontier models — rewards scale and cheap power, and is consolidating into 100MW-to-gigawatt campuses on rural and exurban land. **Inference** — the work of running those models in production — does the opposite. It is being pulled back toward population centers by latency, cost, data gravity, and sovereignty law. This is the central dynamic of the next decade of compute, and the industry has already named it: "Training built the campuses. Inference will choose the markets." (DCD, 2025).

The numbers behind that line are decisive. Inference compute is projected to grow from 20.9 GW in 2025 to 93.3 GW by 2030 — a 35% CAGR — versus training's 23.1 to 62.2 GW at 22% (McKinsey, Jul 2025). By 2030, inference becomes the majority of all AI compute and 30-40% of total data center demand. Deloitte reads the shift faster still, placing inference at roughly two-thirds of AI compute in 2026 and overtaking training in revenue that same year (Deloitte TMT Predictions 2026). And inference, unlike training, has to live near its users: real-time agents, interactive applications, and autonomous systems demand sub-100ms round-trips that are only achievable close to demand.

This is happening on top of a buildout already running at extraordinary velocity. Worldwide data center capital expenditure surged 57% in 2025, with full-year 2026 capex forecast to surpass $1 trillion and grow at a 21% CAGR through 2029 (Dell'Oro Group, Mar 2026). McKinsey frames the cumulative requirement at $6.7 trillion by 2030, of which $5.2 trillion is AI-specific (McKinsey, Apr 2025). The colocation segment — the home of in-city, multi-tenant capacity — is roughly $76-77B globally in 2025, on track for ~$256B by 2035 at an 11.8% CAGR (Expert Market Research; MarketsandMarkets).

The structural opening

As inference demand moves into the metros, the supply that should serve it is moving out.

The listed colocation majors are racing up-market into gigawatt hyperscale, and in doing so are structurally vacating the sub-40MW, in-city tier — exactly the footprint that urban inference needs. Hyperscale already controls ~48% of operational capacity and is forecast to reach 67% by 2031 (Synergy Research, Q1 2026). The economics are pulling every large operator toward the same rural megacampuses, leaving the smaller urban tier underserved.

Meanwhile, the urban supply that exists is locked. North American primary-market vacancy sits near 1.6%, dropping below 1% in Northern Virginia, New York, and Atlanta; Europe's FLAP-D markets hit a record-low 6.3% vacancy with the pipeline 83% pre-let (CBRE, H2 2025; JLL, 2026 Outlook). You cannot rent what does not exist.

The binding constraint underneath all of it is power — and power favors the small footprint. Grid interconnection queues now run four to seven years in Northern Virginia and PJM, five to six-plus in New York, and West London has signaled "no capacity until 2035" (RMI, 2025; DCD). A gigawatt campus must wait in that queue. A small urban node often does not: it can draw on existing building service or behind-the-meter generation, sidestepping the wait entirely. The growing behind-the-meter footprint — tracked at roughly 90 GW — reflects how operators are now routing around the grid where it cannot deliver (Cleanview).

Regulation reinforces the same boundary. The harshest rules are calibrated against giants: large-load tariffs triggered at 25, 75, and 100MW thresholds, the Dutch ban on hyperscale facilities above 70MW, and curtailment

regimes all sit above a modular footprint. Megawatt thresholds, in other words, shape exactly where small footprints retain an advantage (MultiState, 2026; DLA Piper).

Sovereignty is the accelerant. The number of countries with data-protection laws has grown from 76 to over 120 since 2011, with 60-plus now imposing localization rules. Gartner projects that more than 75% of EU and MEA enterprises will move workloads to sovereign solutions by 2030, up from under 5% in 2025 (Gartner, via NetApp). Data that must stay in-jurisdiction forces compute into specific cities and countries — precisely where a small in-city footprint fits.

The under-measured tier

The segment where these forces converge — urban capacity below 40MW — is the one no published market report sizes cleanly. Estimates of the "edge data center" market alone range from $34.8B to $50.9B in 2025 purely on definition, with CAGRs from 14.9% to 22.9% (Grand View; GMInsights; MarketsandMarkets). The directly relevant modular segment is growing from roughly $29-32B to $76-85B by 2030 at a ~17.4% CAGR, with prefabrication cutting time-to-operational by about half (Grand View; Mordor). The urban band straddles the "edge" and "colocation" taxonomies and is under-counted by every published market report — which is both the measure of the opportunity and the reason it has been overlooked.

The technology is moving in the same direction as the demand. The industry is shifting toward bare-metal GPU provisioning, disaggregated inference serving — splitting the phases of model execution across heterogeneous, increasingly air-cooled silicon — and confidential computing that lets tenants trust shared infrastructure. Alongside it sits a second urban workload: private post-training pods, where enterprises fine-tune models on data that cannot leave their control. And increasingly, distributed cross-site inference treats a network of small urban nodes as one fabric. The physical layer to run all of this — urban land, deliverable power, fast modular and brownfield builds, and the cooling and waste-heat reuse that make in-city operation viable — is the scarce input.

BMETAL's role

BMETAL builds the physical layer of AI — urban AI-compute infrastructure in the cities where intelligence is made.

The company secures urban land and power, including behind-the-meter generation where the grid cannot deliver, deploys modular and brownfield compute nodes on fast timelines onto infrastructure-grade real estate, and operates the compute itself — bare-metal GPU capacity and inference — while reusing waste heat. It is built for the moment the market is entering: demand structurally relocating into cities, supply locked, incumbents leaving the sub-40MW tier, power favoring the small footprint, and regulation calibrated against the giants. The physical layer of AI is being built where intelligence is consumed. BMETAL is building it.

The Structural Shift: Inference Moves Into the City

For a decade, the geography of computing was decided by the cost of electricity and land. The next decade will be decided by distance to the user. AI compute is bifurcating into two distinct businesses with two opposite location logics — and the second of these is only now beginning to reshape the map.

Two Workloads, Two Geographies

Training and inference are no longer variations on a theme; they are diverging industries. Training rewards raw scale and the cheapest available power, which pushes it toward 100MW-to-gigawatt campuses on rural and exurban land — the megabuilds rising at Abilene, the Stargate footprint, the Aligned and CoreWeave estates. Inference rewards proximity to where intelligence is actually consumed, which pulls it back toward population centers.

The growth math makes the divergence concrete. Inference capacity rises from **20.9 GW in 2025 to 93.3 GW in 2030, a 35% CAGR**, while training grows from **23.1 GW to 62.2 GW at 22%** (McKinsey, 2025). By 2030, inference becomes the majority of AI compute — over 50% — and **30-40% of total data center demand**. Deloitte reads the crossover as nearer still: inference was roughly **half of AI compute in 2025, rising to two-thirds in 2026**, overtaking training in revenue terms within the year (Deloitte, 2026).

WORKLOAD	2025 CAPACITY	2030 CAPACITY	CAGR	LOCATION LOGIC
Training	23.1 GW	62.2 GW	22%	Scale + cheap power → rural/exurban gigawatt campuses
Inference	20.9 GW	93.3 GW	35%	Proximity → distributes toward population centers

The industry's own shorthand captures the shift exactly: **"Training built the campuses. Inference will choose the markets."** (DCD, 2025). The first phase of the AI buildout was a race for power and land at any distance. The second is a race for position.

The Scale of the Underlying Build

This is not a niche reordering inside a flat market. Worldwide data center capital expenditure **surged 57% in 2025**, with **2026 capex forecast to surpass $1 trillion** and to compound at a **21% CAGR through 2029** (Dell'Oro Group, 2026). McKinsey frames the cumulative requirement as **$6.7 trillion by 2030, of which $5.2 trillion is AI-specific** (McKinsey, 2025). Global capacity is projected to roughly double, adding **~97 GW between 2025 and 2030 to reach ~200 GW** (JLL, 2026); McKinsey's higher-demand case reaches **219 GW by 2030, ~156 GW of it AI**.

The structural shift toward inference, in other words, is happening inside the largest infrastructure buildout of the era. The question is not whether the demand exists — it is where it lands.

Why Inference Pulls Compute Into the City

Four forces, all of them economic rather than aesthetic, draw inference back toward population centers.

- **Latency.** Real-time AI — agents, interactive assistants, AR/VR, autonomous and trading systems — needs sub-100ms round-trips. That budget is only achievable close to the user. A gigawatt campus three states away cannot serve a conversation that must answer in a heartbeat.
- **Cost.** Inference is recurring and revenue-linked, unlike the lumpy capital event of training a model. That changes the calculus on siting: optimizing the per-token cost of a workload that runs forever matters far more than optimizing a one-time build. On-premise and edge inference can deliver **50%+ savings over three years and up to an 18x per-token cost advantage versus cloud APIs** (Deloitte, 2026).
- **Data gravity.** The volume of tokens being processed is exploding, and the data they touch increasingly lives close to the user. As one demand proxy, Google's monthly token processing rose roughly **7x year-on-year** through 2025-2026, and the population of enterprises generating more than 100 billion tokens a month is set to triple between 2025 and 2028 (Tomasz Tunguz). Moving the compute to the data is cheaper than moving the data to the compute.
- **Sovereignty.** Data that must remain in-jurisdiction forces compute in-country, and often in-city. The number of countries with data-protection laws grew from **76 to 120+ between 2011 and 2025**, with 60+ now imposing localization rules. Gartner projects that **more than 75% of EU and MEA enterprises will move workloads to sovereign solutions by 2030, up from under 5% in 2025** (Gartner, via NetApp).

A structural tailwind sits beneath all four: enterprise repatriation. In 2025, **86% of CIOs planned to bring at least some workloads back** from public cloud — the highest reading on record — driven by AI, cost, and sovereignty (IDC). The center of gravity is moving toward owned, proximate, in-jurisdiction compute.

The Segment the Shift Creates

As inference distributes, demand concentrates in a tier of the market that incumbents are actively vacating. Hyperscale already controls **~48% of operational capacity** and is climbing toward **67% by 2031** as operators race up-market to gigawatt campuses (Synergy Research, 2026). That ascent leaves the sub-40MW urban footprint — the band where proximity-sensitive inference must live — structurally underserved.

The market that serves it is growing fast and is, tellingly, hard to measure. Colocation stands at roughly **$76-77B globally in 2025, heading to ~$256B by 2035 at an ~11.8% CAGR**, with wholesale colocation the fastest sub-segment at ~17% (Expert Market Research / MarketsandMarkets). The edge layer is harder to bound: estimates of the 2025 edge data center market range from **$34.8B to $50.9B** depending purely on definition, with **micro data centers the fastest-growing size class at ~35% of the edge market** (Grand View Research). The directly relevant modular segment runs **~$29-32B in 2024 to $76-85B by 2030 at ~17.4% CAGR**, with prefabricated builds cutting time-to-operational by roughly half — the enabler of fast deployment into constrained urban sites (Grand View / Mordor).

The urban band straddling near-edge and small colocation is precisely the slice most published taxonomies under-count, because it falls between the "edge" and "colocation" categories. That measurement gap is the clearest signal of how new the opportunity is — the demand is arriving faster than the analysts have agreed on what to call it.

The Takeaway

Training built the campuses. It optimized for power and land at distance, and it produced an extraordinary concentration of compute far from where most intelligence is consumed. Inference inverts every term of that equation: it optimizes for proximity, runs on recurring revenue, and answers to latency, cost, data gravity, and sovereignty — all of which point in the same direction, toward the city. The physical layer of AI is being pulled into the metros, and the market structure to serve it is only beginning to form.

Market Size & Demand

The defining feature of the AI-compute market is no longer its size — it is its velocity. Worldwide data center capital expenditure surged 57% in 2025, and full-year 2026 capex is forecast to surpass $1 trillion, compounding at a 21% CAGR through 2029 (Dell'Oro Group, 2026). McKinsey frames the cumulative buildout requirement at $6.7 trillion by 2030, of which $5.2 trillion is AI-specific (McKinsey, 2025). This is a structural reallocation of global infrastructure spend, not a forecast waiting on demand to materialize.

By revenue, the global data center market stood at roughly $384–418B in 2025, on a trajectory to about $692B by 2030 (~10.6% CAGR) on the Grand View / Precedence consensus (Grand View Research, 2025; Precedence Research, 2025). Physical capacity is set to roughly double — adding about 97 GW between 2025 and 2030 to reach ~200 GW (JLL, 2026) — with McKinsey's higher-demand case reaching 219 GW by 2030, of which ~156 GW (~70%) is AI.

Segment structure: hyperscale, colocation, edge

The market resolves into three tiers with sharply diverging dynamics.

SEGMENT	2025 SIZE	TRAJECTORY	SOURCE
Hyperscale	~48% of operational capacity (Q4 2025), ~1,360 facilities	Rising to 67% by 2031; capacity roughly triples by 2030	Synergy Research, 2026
Colocation	~$76–77B global	~$256B by 2035 (~11.8% CAGR); wholesale colo fastest at ~17% CAGR	Expert Market Research, 2025; MarketsandMarkets, 2025
Edge	~$34.8–50.9B (range driven by definition)	14.9%–22.9% CAGR; micro DCs ~35% of edge and the fastest size class	Grand View Research, 2025

Hyperscale dominates headline capacity and is racing up-market into 100MW-to-gigawatt campuses. Colocation is the largest near-term commercial pool. Edge is the fastest-growing — and the least consistently measured.

Regional spend concentrates in the mature metros and the fastest-rising emerging ones:

REGION	COLOCATION 2025	COLOCATION 2030	CAGR
North America	$38.8B	$65.5B	~11%
Western Europe	$22.0B	$41.1B	~13.3%
Latin America (total DC)	~$16.6–19.8B	~$33B	~11% (colo specifically ~25.7%)

(Sources: Mordor Intelligence, 2025; IMARC, 2025; Arizton, 2025.)

The structural split: training campuses vs inference markets

The most consequential dynamic in the market is the bifurcation of AI compute. Training rewards scale and cheap power, consolidating into 100MW-to-gigawatt campuses on rural and exurban land. Inference rewards proximity, distributing compute back toward population centers.

The trajectories diverge accordingly. Inference compute rises from 20.9 GW in 2025 to 93.3 GW by 2030 — a 35% CAGR — versus training's 23.1 to 62.2 GW at 22% (McKinsey, 2025). By 2030, inference exceeds 50% of AI compute and accounts for 30–40% of total data center demand. Deloitte reads the shift as faster still: inference was roughly 50% of AI compute in 2025, rising to two-thirds in 2026 and overtaking training in revenue during 2026 (Deloitte, 2026).

Three forces pull inference into cities:

- **Latency.** Real-time AI — agents, AR/VR, autonomous systems, interactive applications — requires sub-100ms round-trips, achievable only near end users.
- **Cost.** Inference is recurring and revenue-linked, unlike lumpy training capex. On-premises and edge inference can deliver 50%+ savings over three years and up to an 18× per-token cost advantage versus cloud APIs (Deloitte, 2026).
- **Data gravity and sovereignty.** Data that must remain in-jurisdiction forces in-country — and frequently in-city — compute.

As the industry puts it: "Training built the campuses; inference will choose the markets" (DCD, 2025). Underlying demand is visible in token throughput: one major platform's monthly processing climbed from ~9.7 trillion tokens (April 2024) toward the quadrillion scale by 2026 — roughly 7× year-on-year (vendor self-reported; directional). Enterprises generating more than 100 billion tokens per month are projected to triple between 2025 and 2028 (Tomasz Tunguz).

The under-counted urban band

The segment with the clearest structural tailwind is also the worst measured. Estimates of the edge data center market range from $34.8B to $50.9B in 2025 — a nearly 3× spread driven purely by definition (Grand View ~$34.8B at 14.9% CAGR; GMInsights ~$50.9B; MarketsandMarkets $22.2B on a narrower scope) (Grand View Research, 2025). Micro data centers represent roughly 35% of the edge market and its fastest-growing size class.

The directly adjacent modular segment runs from ~$29–32B (2024) to $76–85B (2030) at a ~17.4% CAGR, with prefabricated builds at ~63% of the total and containerized modules growing ~19.5% (Grand View Research, 2025; Mordor Intelligence, 2025). Prefabrication cuts time-to-operational by roughly 50% and on-site work by up to 60% — the core enabler of fast urban deployment.

The definitions matter because the most attractive band falls between them:

- **Edge DC:** typically 500 kW–2 MW, at the network's outer edge.
- **Micro DC:** a pre-integrated unit, ~100–150 kW critical load, often containerized.
- Analysts further split **"near edge"** (metro/regional, 1–10 MW) from **"far edge"** (on-premises/cell-site, sub-MW).

The sub-40MW urban tier — small in-city micro footprints plus mid-sized metro colocation — straddles "near edge" and "small colocation." It is the slice most published reports under-count, because it sits between the edge

and colocation taxonomies rather than inside either. That structural under-measurement is precisely where the supply-demand mismatch is sharpest: incumbents are vacating the sub-40MW urban tier as they scale to gigawatt campuses, even as inference pulls recurring-revenue compute into the same metros.

The demand drivers, weighted

1. **AI inference latency (primary).** The structural catalyst — it pulls recurring-revenue compute into metros by economic necessity, not preference.
2. **Real-time and interactive applications.** Gaming, AR/VR, autonomous systems, and latency-sensitive finance all require sub-100ms inference. Not separately sized in credible 2025 reports; best characterized as latency-driven pull demand.
3. **Data sovereignty and residency.** Countries with data-protection laws grew from 76 to 120+ between 2011 and 2025, and 60+ now impose localization rules. Gartner projects that 75%+ of EU and MEA enterprises will geopatriate workloads to sovereign solutions by 2030, up from under 5% in 2025 (NetApp / Gartner, 2025). This forces compute into specific cities and countries.
4. **Telco / 5G / MEC edge.** Multi-access Edge Computing colocates compute at the radio and metro edge — a primary driver behind Latin America's ~25.7% colocation CAGR and global edge growth.

A structural tailwind reinforces all four: enterprise repatriation. In 2025, 86% of CIOs planned to repatriate at least some workloads — the highest share on record — driven by AI, cost, and sovereignty (IDC, 2025).

Supply is the binding constraint

The binding constraint on the demand side is not appetite but supply. The inference shift — reaching 30–40% of total data center demand by 2030 (McKinsey, 2025) — intersects record-low metro vacancy: CBRE recorded 0.5% vacancy in Northern Virginia, and Cushman & Wakefield 6.3% across the core European FLAP-D markets (CBRE, 2025). Demand is structural and accelerating; in the cities where intelligence is made, supply is locked.

> **A note on the figures.** Edge-segment 2025 sizing ranges ~3× by definition — no single number should be cited without its source's scope. Latin American headline market size varies ~17% across firms. Token-volume figures are vendor self-reported and should be read as directional.

The Supply Crunch: Why Incumbents Can't Follow

The urban AI-compute thesis rests on a deceptively simple fact: capacity that does not exist cannot be leased. Demand for in-city inference is rising on the back of latency, cost, data gravity, and sovereignty — but the supply that could meet it is structurally blocked at every layer. Primary-market vacancy has collapsed to record lows. Grid interconnection has become a multi-year gating event. Urban land that can carry a data-center load is scarce and contested. And the colocation majors who once built mid-sized in-city facilities are racing up-market to gigawatt campuses, vacating the sub-40MW urban tier precisely as demand returns to it. The result is a widening gap between where intelligence needs to be served and where capacity is physically available.

Vacancy has collapsed in the primary markets

The first signal of a supply crunch is price and availability, and both point the same way. In the primary North American markets, vacancy has fallen to levels with no historical precedent. Northern Virginia — the largest data-center market on earth — sits at **0.5% vacancy** (CBRE, 2025). Across the major European FLAP-D markets (Frankfurt, London, Amsterdam, Paris, Dublin), vacancy is roughly **6.3%** (Cushman & Wakefield, 2025). These are not soft markets correcting toward equilibrium; they are markets where, in practical terms, there is nothing left to lease.

Scarcity is also visible in pricing. High-density wholesale capacity in the 250–500 kW band reached a record **$196.25/kW/month** (CBRE, 2025) — a clearing price that reflects acute supply tightness rather than ordinary inflation. When the most contested capacity prices like a scarce commodity, it is because it is one.

The grid is the binding constraint

Behind the vacancy numbers sits the true bottleneck: power. Securing a grid connection for a new load has gone from a procedural step to a multi-year gating event, and the wait times in the markets where urban demand concentrates are now measured in years.

MARKET	INTERCONNECTION REALITY
Northern Virginia / PJM	Dominion warns connection times of up to **7 years** (from 3–4); PJM application-to-operation has risen from under 2 years (2008) to **over 8 years (2025)**
New York / Con Edison	NYISO large-load queue reached **51 projects / 12,670 MW (May 2026)**, against historical 5–6+ year waits
Dublin / Ireland	De facto moratorium lifted Dec 2025, but new data centers must self-generate full demand and source at least 80% from new renewables
Amsterdam / Netherlands	Over **12,000 firms** waiting on grid connections; Amsterdam reportedly closed to new data-center applications
West London	**"No spare capacity until 2035"** — severe enough to stall housing

(Sources: RMI, 2025; DCD on Dominion; CRU, 2025; NL Times, 2025; DCD, 2024.)

The scarcity is now quantified in the markets themselves. In December 2025, PJM's capacity auction cleared **6,625 MW short** of target for the first time in its history, with capacity prices leaping from $28.92/MW-day (2024/25) to **$333.44/MW-day** — roughly 40% of that auction's cost attributed to data centers (Utility Dive, 2025; PJM Inside Lines, 2025). When the grid clears short and prices multiply by more than ten, the message to a would-be large load is unambiguous: there is no power to connect, and no near-term path to it.

There is a structural asymmetry hidden in these figures that is decisive for the urban tier. Every one of these constraints is calibrated to large loads. The interconnection queues that paralyze gigawatt campuses are the binding constraint for hyperscale; a sub-5MW node can frequently draw on existing building service, a modest upgrade, or behind-the-meter generation, sidestepping the queue entirely. The same megawatt thresholds that shape large-load tariffs and moratoria — Ireland's self-generation mandate, Amsterdam's closure, large-load capacity charges in PJM — are written for the loads that small in-city footprints fall beneath. Speed-to-power, not cheap power, has become the defining constraint of the market — and the megawatt thresholds embedded in tariffs and moratoria fall hardest on large loads, leaving small in-city footprints structurally advantaged on time-to-energize.

Urban land that can carry the load is scarce

Power is the first constraint; the building is the second. Data centers require roughly **150–250+ lbs/sq ft of floor loading** (versus ~60–100 for offices) and over **30× the watts per square foot** of conventional commercial space (brownfield-conversion analysis, HKS / Bisnow, 2025). That eliminates most of the urban building stock outright: offices and multi-story retail are the hardest, least-viable conversions. The viable urban shells are specific and limited — telco central offices (built with 48V DC power, multi-hour battery, existing fiber, and ~30ms latency to users) and heavy industrial or warehouse buildings (with the floor load, clear span, and power that data centers need).

These assets are finite, and the field has noticed. Telco operators are already converting their estates — Ziply is converting 200+ exchanges; BT is retaining roughly 1,000 — which both proves the model and signals that the most attractive inherited-power shells are being claimed. Where existing power, structure, and fiber can be reused, retrofit reaches stabilization in **12–24 months** versus 24–48 for greenfield. The constraint is not whether brownfield works; it is that the supply of urban shells with inherited power is scarce and shrinking, even as the demand pulling compute back into cities accelerates.

The incumbents are structurally vacating the tier

The final, and most important, reason supply cannot meet urban demand is that the players best positioned to build it are deliberately moving the other way. The listed colocation and wholesale majors are racing up-market into gigawatt hyperscale on cheap exurban land — not down into the dense, sub-40MW urban tier.

- **Digital Realty** is bringing roughly **5 GW** online across 40 metros and has explicitly described "doubling down on largest markets" — the opposite direction from in-city.
- **Equinix** is targeting a capacity doubling by 2029, with its xScale hyperscale unit leasing 430 MW; its strategic energy lies in interconnection fabric, not small-urban builds.
- **CyrusOne, Vantage, NTT, and QTS** are all hyperscale and wholesale builders — CyrusOne alone has roughly 674 MW in operation plus a 760 MW Texas campus.
- **Iron Mountain**, the most relevant brownfield-retrofit analog, is tripling its data-center business to **1.3 GW** — proving brownfield conversion is bankable at scale, but doing it in service of large-format capacity.

(Sources: DCD, 2025; Bisnow, 2025; NTT DATA / ABI Research; Iron Mountain 8-K, SEC.)

This is not a temporary tactical preference; it is dictated by the economics of the buildout. With worldwide data-center capital expenditure having surged **57% in 2025** and 2026 capex forecast to surpass **$1 trillion** (Dell'Oro Group, 2026), and with McKinsey framing the cumulative requirement at **$6.7 trillion by 2030** (McKinsey, 2025), the rational move for a capital-intensive operator is to chase the largest contiguous loads on the cheapest power. Training rewards scale and cheap power, consolidating into 100MW-to-GW campuses on exurban land. The majors are following that gradient up and out — which is exactly why the in-city tier is being left behind.

A cautionary history reinforces why the gap persists. The previous wave of edge-native operators learned that meaningful metro-interconnect at scale thrives, while ultra-distributed micro-edge fails. The winners — DataBank (65+ facilities across 25+ metros), Cologix (45+ facilities across 13 markets), EdgeConneX (which partnered with a neocloud on 30+ MW of high-density AI in Chicago) — built dozens of 5–40MW metro nodes. The "tiny boxes everywhere" thesis collapsed: Vapor IO, which raised over $200M to put micro-edge in 50–500 metros, moved away from the model entirely; American Tower deployed a handful of tower-base edge sites, then retreated and acquired CoreSite ($10.1B) instead; Crown Castle's CEO stated flatly, "we're not interested in investing in data centers." The lesson is that the viable urban node size is real and proven — and that the incumbents who could serve it have either moved up-market or exited.

The arithmetic of the gap

Put the pieces together and the conclusion is structural, not cyclical:

- **Demand is moving into cities.** Inference rises from **20.9 GW (2025) to 93.3 GW (2030)** at a 35% CAGR, overtaking training and reaching more than half of AI compute by 2030 (McKinsey, 2025); Deloitte reads inference at roughly two-thirds of AI compute in 2026 (Deloitte TMT Predictions 2026). Inference rewards proximity, and proximity means metros.
- **Primary-market supply is exhausted.** Vacancy of 0.5% in Northern Virginia and ~6.3% across FLAP-D leaves no slack to absorb that demand (CBRE, 2025; Cushman & Wakefield, 2025).
- **New supply is grid-blocked.** Interconnection waits of 7–8+ years, a PJM auction clearing 6,625 MW short, and outright moratoria mean large-format capacity cannot be added on any timeline that matches demand (RMI, 2025).
- **The incumbents are leaving.** The colocation majors are doubling down on gigawatt exurban campuses, and no major operator is optimizing for the dense, sub-40MW urban tier.

The intersection of these forces — inference reaching 30–40% of total data-center demand by 2030 set against record-low metro vacancy — is the binding fact of the market: demand exists, and supply is locked. Capacity that doesn't exist cannot be leased. The urban tier is being structurally vacated at the precise moment the economics of AI are pulling intelligence back into the cities where it is made. That is the supply crunch — and it is why the incumbents, by their own design, cannot follow.

Power: The Binding Constraint

Across the world's leading metros, the limiting factor for AI compute is no longer capital, land, or even chips. It is power — specifically, the ability to secure a grid connection on a timeline that matches the pace of AI demand. Access, not price, has become the constraint. Understanding why is essential to understanding where AI compute can actually be built, and why small urban footprints can move while gigawatt campuses wait.

The interconnection queue has become a multi-year gating event

Securing a grid connection for a new large load has shifted from a procedural step into one of the longest-lead items in the entire build. In the largest US market, Dominion has warned that connection times in Northern Virginia can run as long as seven years, up from three to four (DCD). Across PJM, the time from application to operation has stretched from under two years in 2008 to more than eight years by 2025 (RMI). The strain is now showing up in the market itself: PJM's December 2025 capacity auction cleared 6,625 MW short of target for the first time ever, and capacity prices leapt from $28.92/MW-day in 2024/25 to $333.44/MW-day, with roughly 40% of that auction's cost attributed to data centers (Utility Dive; PJM Inside Lines).

The pattern repeats across every major demand center:

MARKET	INTERCONNECTION REALITY
Northern Virginia / PJM	Connection times up to 7 years (from 3–4); PJM application-to-operation now >8 years; Dec 2025 capacity auction cleared 6,625 MW short for the first time
New York / Con Edison	NYISO large-load queue reached 51 projects / 12,670 MW (May 2026); historical 5–6+ year waits
Dublin / Ireland	De facto moratorium lifted Dec 2025, but new data centers must self-generate full demand and source ≥80% from new renewables; data centers already ~21% of national electricity
Amsterdam / Netherlands	>12,000 firms waiting on grid connections ("netcongestie"); Amsterdam effectively closed to new applications
West London	"No spare capacity until 2035" — severe enough to stall housing
Querétaro / Mexico	CFE timeline ~14 months just to transformer purchase; 540 MW in outstanding requests
Santiago / Chile	Data center demand forecast +270% (325→1,207 MW, 2025–30) against a constrained grid

Sources: RMI; DCD; CRU; NL Times; Mexico Business; Energy Partnership Chile.

The structural point: nearly every one of these constraints is calibrated to large loads. The thresholds, queues, and tariffs that paralyze gigawatt-scale interconnection are precisely the thresholds a small urban footprint can fall beneath. Small footprints can frequently draw on existing building service, a modest service upgrade, or on-site generation — sidestepping the queue entirely. In a market defined by speed-to-power, megawatt thresholds shape where small footprints hold a structural advantage.

Power economics: a 4× spread, and demand charges that punish lumpy load

Where power is available, its price varies nearly fourfold across leading metros — enough to be a primary siting driver in its own right.

MARKET	LARGE-LOAD PRICE (2025–26)	USD/KWH
US average	8.6¢	$0.086
Virginia	10.25¢ (+15% YoY)	$0.103
New York (industrial)	8.73¢	$0.087
Mexico (GDMTH)	—	~$0.12
Brazil / São Paulo	—	~$0.16
Chile / Santiago	—	~$0.20
Ireland (large band)	19.1 €¢	~$0.21
UK (very-large industrial)	22.4p	~$0.30

Sources: EIA (Mar 2026); UK DESNZ (Sep 2025); Eurostat / Statista.

Two nuances reshape the conventional wisdom. First, New York's reputation for expensive power is largely a commercial-tariff and delivery-charge phenomenon (~22¢), not a reflection of statewide industrial rates, which sit near 8.7¢. Second, the UK and Ireland are becoming cost-prohibitive on price alone, while Spain ranks among the cheapest in Western Europe — a divergence that increasingly governs where European AI compute concentrates.

Beyond the headline rate, demand charges are rising sharply and can account for 30–70% of a demand-metered bill. This has turned on-site battery storage from a luxury into a standard urban add-on: behind-the-meter storage for demand-charge management is now a baseline design element across the industry, not a differentiator.

Behind-the-meter generation: "bring your own power"

The defining infrastructure trend of 2025–26 is "bring your own power" (BYOP) — and it is driven by time-to-power, not cost. Roughly 90 GW of behind-the-meter capacity has been announced, representing more than a quarter of planned US data center capacity, and 92% of it has been announced since January 2025 (Cleanview). When the grid cannot deliver on a workable timeline, on-site generation becomes the path of least resistance.

At urban scale, the realistic generation stack is narrow but proven:

- **Solid-oxide fuel cells are the standout urban option.** Running on natural gas, biogas, or hydrogen with no combustion, they reach roughly 60% electrical efficiency (90% in combined heat and power), deploy in as little as 55–90 days, and produce near-zero NOx and SOx — often making them air-permit-exempt. At around 65 dB and roughly 30 MW/acre of power density, they fit downtown footprints where turbines cannot. The urban-colocation precedent already exists at scale: Bloom Energy has signed deals with Oracle (up to 2.8

GW), AEP (up to 1 GW), and — most tellingly — Equinix, which has deployed more than 100 MW across roughly 19 urban colocation sites (Bloom Energy).

- **Lithium battery storage is the standard companion** — for demand-charge management (30–60% reductions), peak-shaving, and ride-through. Pricing has fallen to roughly $110/kWh globally (~$219 in the US) for four-hour systems (Energy-Storage.News). Indoor deployment is capped by NFPA 855 fire-area limits.

Several options that dominate the hyperscale conversation are not viable at urban scale:

- **Natural gas turbines** carry lead times of 18 months to seven years — GE Vernova reports a 100 GW backlog and Mitsubishi is sold out into 2028 — and belong to the gigawatt tier (Power Engineering; Utility Dive).
- **Solar** is mathematically impossible at urban density, requiring 7–10 acres per MW; a fully covered roof supplies less than 5% of load (Scale Microgrids).
- **Small modular reactors** remain a hyperscale-and-2030+ proposition; no SMR product exists for urban-scale deployment on any realistic timeline.
- **Combustion generators** face year-plus air permits, residential noise limits of roughly 45–55 dBA, and gas-line constraints that frequently make them impractical downtown.

The practical urban stack, then, is the grid as primary supply, solid-oxide fuel cells as the best on-site generation option, lithium storage for peak-shaving and ride-through, and code-required emergency backup only.

Rising rack density favors the urban footprint

A counterintuitive dynamic works in favor of land-constrained urban sites: as rack density climbs, each square meter of floor becomes more productive.

Traditional enterprise racks ran 5–10 kW; the 2025 installed base averages roughly 7.6–9 kW, with new builds near 17 kW and trending toward 30 kW by 2027. AI accelerates this sharply — NVIDIA's GB200 NVL72 draws roughly 120 kW nominal and 130–140 kW as deployed, with mandatory liquid cooling, and the public roadmap points toward 600 kW racks by 2027 and eventually toward 1 MW racks on 800 VDC architectures (Introl; Spheron; DCD). Around 50 kW per rack is the air-cooling cliff; above roughly 70 kW, liquid is the only practical option.

The footprint implication is direct. AI facilities draw 300+ W per square foot versus 100–150 for conventional builds — roughly 3× the productive compute per square foot — which lets capacity scale within an existing footprint rather than expand it. Density also commands a rent premium: 250–500 kW wholesale capacity reached a record $196.25/kW/month (CBRE). Where urban floor area is scarce and expensive, high density is not a cost problem to be managed but a structural advantage — and increasingly the industry default is to design liquid-cooled from the outset.

The other density story: air-cooled inference silicon

The rack-density narrative has two halves, and most coverage tracks only one. The hyperscale training tier is the one that gets the headlines: NVIDIA's GB200/GB300 NVL72 racks sit at roughly 120–140 kW today, with the public roadmap pointing toward 600 kW and eventually 1 MW racks — densities that make liquid cooling mandatory and force a gut-renovation of any building's HVAC, electrical, and water systems. But a parallel and fast-emerging class of inference-optimized accelerators occupies the opposite end of the envelope: on the order of 8–12 kW per 6U system, air-cooled, and scaling over commodity 800GbE Ethernet rather than proprietary fabric — for example Tenstorrent's Galaxy systems, with entry pricing reported around $110k per unit (vendor

figures pending independent third-party benchmarks; Tenstorrent; HPCwire). For the urban thesis this distinction is decisive. Air-cooled inference silicon fits within the power-and-cooling envelope of a converted Class-B urban building that could never host a liquid-cooled hyperscale rack, lowering the retrofit bar and widening the set of in-city shells that can carry meaningful inference capacity.

Cooling: liquid is now the default, and it favors cities

The economics of cooling have crossed a threshold. The total-cost crossover between air and liquid cooling arrives at roughly 30 kW/rack at $0.12/kWh — well below where AI workloads now sit. Direct-to-chip dominates AI liquid cooling at roughly 47% share, with immersion serving the highest densities. More than half of new hyperscale capacity is expected to be liquid-cooled by 2027, and the liquid-cooling market is projected to grow from roughly $5B in 2025 to ~$7B by 2029 (Dell'Oro).

Crucially for urban siting, liquid cooling uses roughly 60% less floor space than air, even though facility capex runs higher (~$3–4M/MW versus ~$1.5–2M/MW for air). It also resolves an emerging permitting risk. Global average PUE has stagnated at 1.54 for six years (Uptime Institute), but closed-loop liquid systems reach 1.05–1.2 with near-zero water consumption. With more than 40% of US data centers located in high-water-stress areas and roughly 30 state water bills introduced in 2026 (Nixon Peabody; MultiState), closed-loop liquid cooling neutralizes the water objection that increasingly drives local opposition.

Waste-heat reuse: from disposal cost to revenue line

In Europe especially, the heat a data center rejects is becoming both a revenue stream and a license to operate. Stockholm Exergi pays roughly 2M SEK per year for each MW of delivered heat — one MW serving the equivalent of about 1,500 apartments (Stockholm Exergi). Policy is now mandating the practice: Germany's Energy Efficiency Act (EnEfG) legally requires waste-heat reuse, with energy-reuse factors rising on a fixed schedule (≥10% in July 2026, ≥15% in 2027, ≥20% in 2028) plus a PUE ceiling of 1.2 for new builds above 300 kW; France requires an energy-reuse factor of ≥0.2 for facilities above 1 MW from January 2026, with penalties up to €50k per site (White & Case). Flagship projects already demonstrate the model at scale — Meta's Odense facility delivers 215,000 MWh/yr to more than 12,000 homes, and the Microsoft/Fortum project in Finland is set to serve roughly 250,000 customers, the world's largest waste-heat recycling effort (Fortum). For an in-city European facility within roughly 5 km of a district-heating network, a heat offtake is viable at $15–30/MWh-thermal — converting waste heat into both social license and a revenue line.

Modular and brownfield: building around the procurement bottleneck

If the grid is one bottleneck, equipment procurement is the other. Transformers now carry lead times of roughly 18 months and switchgear roughly 22 months — and on a large facility, a single month of delay can erase millions in foregone revenue. Two build methods directly attack this constraint.

Modular and prefabricated construction compresses lead times to 2–12 months versus 18–36 for conventional builds, while delivering cost certainty, factory quality, and revenue roughly three months sooner. A robust vendor ecosystem now supports high-density, liquid-ready prefab:

VENDOR	PRODUCT LINE	NOTE
Vertiv	SmartMod / MegaMod CoolChip	High-density AI prefab, direct-to-chip liquid
Schneider / EcoStruxure	Prefab Modular Pod	40+ high-density racks, hybrid liquid-air

VENDOR	PRODUCT LINE	NOTE
Huawei	FusionModule / FusionDC	"50%+ delivery-time reduction"
Eaton	Modular MDC + Flexnode	High-density halls, 800 VDC
Dell	Modular Data Center / DC-MHS	Edge line
Armada	Galleon / Leviathan	Containerized to MW-scale, weeks to deploy

Source: vendor materials.

Brownfield retrofit is resurgent because AI edge deployments — measured in dozens of MW rather than hundreds — fit existing shells. The most valuable targets are telco central offices, which arrive with 48V DC power, 48+ hours of battery, existing fiber, and roughly 30 ms latency to users: the costliest elements pre-built. Carriers are already converting their estates — Ziply is converting 200+ exchanges and BT is retaining roughly 1,000 (Bisnow; Cabling Install). Heavy industrial and warehouse shells rank second, offering floor loading, clear spans, and power — exemplified by WhiteFiber/Bit Digital's ~1M sq ft North Carolina factory conversion (Bit Digital).

Structural limits are decisive here: data centers require roughly 150–250+ lbs/sq ft of floor loading (versus 60–100 for offices) and more than 30× the watts per square foot, which makes offices and multi-story retail the hardest, lowest-yield conversions. Where shell, power, and fiber are genuinely reusable, brownfield reaches stabilization in 12–24 months (versus 24–48) at 15–30% lower build cost — and, critically, with inherited power that skips the interconnection queue.

The through-line

Every thread here points the same direction. The interconnection queue is calibrated to large loads, which means small footprints can move while gigawatt campuses wait. On-site generation has shifted from cost play to time-to-power play, and a narrow but proven urban stack — fuel cells plus storage — exists today. Rising rack density makes scarce urban floor area more productive, not less. Closed-loop liquid cooling resolves the water objection, and in Europe waste heat becomes a revenue line. Modular construction and brownfield retrofit route around the procurement bottleneck and, in the case of telco shells, inherit the very power the grid cannot otherwise deliver. Power is the binding constraint of the AI build-out — and the small urban footprint is the form factor built to sidestep it.

Regulation & Policy

Regulation is the most underappreciated force shaping where AI compute will actually be built. The prevailing assumption is that policy is uniformly hostile to data centers. The reality is more precise, and more consequential: almost every binding restriction now on the books is calibrated to giant facilities. Grid moratoriums, large-load tariffs, megawatt-keyed zoning triggers, and the Dutch hyperscale ban are all written against scale. The result is a market structure in which footprint determines regulatory exposure — and in which small, urban deployments occupy a categorically different position than gigawatt campuses.

This is not a loophole. It is the central fact of how policy is redrawing the map of AI infrastructure.

Megawatt thresholds as market structure

The clearest signal of policy direction is the wave of US large-load tariffs — utility rules that impose long contracts, high minimum subscriptions, and exit penalties on the biggest interconnections. What matters is not just that they exist, but where the lines are drawn.

JURISDICTION	THRESHOLD	REPRESENTATIVE TERMS
Ohio (AEP)	>25 MW	≥85% subscribed load, 12-year minimum, exit fee
Virginia (Dominion GS-5)	>25 MW	14-year contracts, ≥85% T&D demand charges (effective Jan 2027)
Georgia	≥100 MW	15-year contracts, upfront cost recovery
Texas (SB 6)	>75 MW	Curtailment / forced load-shed, $100k study fee
Indiana (HB 1007)	Data centers	Cover 80% of new-generation cost even if unbuilt

The pattern is unambiguous. The thresholds that trigger the harshest treatment — 25, 75, 100 MW — sit far above the 1–10 MW band of a modular urban node (MultiState; Utility Dive). The ratepayer backlash now driving these tariffs is a hyperscale problem. Small in-city sites connect under standard commercial tariffs and fall beneath the punitive large-load regime entirely. Megawatt thresholds, in other words, are not a constraint on the urban tier — they are a structural feature that defines its operating space.

Zoning and permitting: the move to industrial

Across US jurisdictions, the dominant 2024–26 trend is reclassification of data centers as industrial uses — confining them to industrial zones and subjecting them to special-use review. Loudoun County eliminated by-right development in March 2025; Fairfax now requires 200-foot residential setbacks and noise studies; Kansas City requires will-serve letters as of January 2026 (Holland & Knight; JLARC; Smart Cities Dive).

The fragility of large-format permitting is best illustrated by the collapse of Northern Virginia's flagship Prince William Digital Gateway, a $35B-plus program voided in August 2025 and abandoned by April 2026 (WTOP; McGuireWoods). It is a vivid lesson in how exposed gigawatt-scale projects have become to local process — and how durable speed-to-market is for footprints that clear permitting without a multi-year fight.

Tiered-by-size frameworks are beginning to emerge — DeKalb County, Georgia, now keys review to square footage and power draw — but no jurisdiction has yet created a distinct lighter track explicitly for small or edge

facilities (EESI). The advantage to small footprints is therefore indirect but real: falling beneath megawatt triggers, plus the inherent speed of modular and brownfield build.

One friction does not shrink with footprint, and it is the most consistent constraint in US zoning: noise. Prince William enforces 60 dB day / 55 dB night; Chandler, Arizona, ties permitting to a multi-year pre-construction acoustic baseline and rejected a data center on those grounds in 2025 (EESI; GPB). Diesel backup compounds the issue — generators run at roughly 105 dB, and proposed Northern Virginia genset fleets were projected to emit on the order of 9,000 tons of NO_x per year. These pressures scale down far less than land and power do, which is why battery and grid-tied backup with acoustic enclosure has become the prevailing design approach for in-city sites rather than diesel.

Europe: thresholds that catch small sites

If US megawatt thresholds favor the urban tier, the EU's efficiency and reporting regime is the counterweight. Here the thresholds are low enough to capture small footprints — and compliance becomes a cost of doing business from the first kilowatt.

- **Energy Efficiency Directive (EED) and Delegated Regulation 2024/1364:** applies at **≥500 kW IT power**, with mandatory annual reporting to the European data center database — energy, renewable share, PUE, WUE, ERF, and waste heat — due by 15 May (EUR-Lex; European Commission). Even small urban colocation is in scope.
- **Germany (EnEfG), the strictest regime in the West:** triggers at **≥300 kW** for colocation. It mandates PUE ≤1.2 for new builds from July 2026, waste-heat reuse rising from **10% to 15% to 20%** across 2026–28, renewables from 50% in 2024 to **100% by January 2027**, plus an energy-management system and regulatory reporting (White & Case). A proposed waste-heat exemption applies only where no district heating exists within 5 km — but an urban operator is almost always within 5 km, so the mandate binds in-city.
- **France:** ≥500 kW reporting, and facilities above 1 MW must externally recover waste heat (ERF ≥0.2) from January 2026, with penalties up to €50k per site.
- **Climate Neutral Data Centre Pact (voluntary, covering >85% of EU capacity):** new-build PUE ≤1.3, 100% carbon-free by 2030, WUE ≤0.4 L/kWh.

The strategic implication runs in both directions. These mandates raise the barrier for everyone, and they make heat reuse and closed-loop water design requirements rather than features. In precisely the markets where the rules bite hardest, compliance from the outset is the price of entry.

Moratoriums giving way to conditional allocation

The most important shift in 2025 is the move from outright prohibition toward conditional, efficiency-gated allocation.

- **Netherlands:** Since January 2024, hyperscale defined as >10 ha **and** >70 MW is banned outside designated zones (DLA Piper; CMS). Small and modular sites fall below the definition, though Amsterdam's municipal grid limits still apply.
- **Ireland:** The de facto Dublin moratorium ended in December 2025, replaced by conditions for new applications — on-site or proximate generation matching maximum import demand, at least 80% from new Irish renewables, and a six-year build window (CRU; Pinsent Masons).

- **Singapore:** Having moved from a 2019 freeze to conditional allocation, its DC-CFA2 round in December 2025 allocated ≥200 MW gated on ≥50% green power and PUE ≤1.25 (Morgan Lewis; IMDA). It is the template other governments now follow.
- **Frankfurt:** The 2022 Masterplan zones the city into suitable, restricted, and exclusion areas, requires green power, roof and façade greening, and district-heat feed-in, and instructs operators to build taller, not wider — with suitable land expected to exhaust around 2025 (DLA Piper; Bird & Bird). Scarcity at scale favors small infill.

The throughline: the rules that constrain hyperscale create the urban gap, while the same policies generate the residency and latency demand that fills it.

Data sovereignty as a demand driver

Regulation does not only restrict where compute can be built — it dictates where it must be. Data localization and sovereignty rules are now an explicit demand engine.

Brazil's Nuvem de Governo sovereign-cloud framework requires classified government data to remain in-country. Mexico's overhauled data regime, effective March 2025, lacks defined cross-border transfer mechanisms, creating de facto pressure toward local and hybrid infrastructure. Chile's GDPR-aligned Law 21.719 takes effect in December 2026 (White & Case; EY). In Europe, sovereign-cloud sentiment — including Dutch parliamentary motions to reduce reliance on US cloud providers — anchors demand in EU-resident compute. Independently, Gartner projects 75% geopatriation of data by 2030 (NetApp), with IDC reporting 86% of CIOs intending some repatriation.

Sovereignty law, combined with latency, produces compute that must be local. It forces workloads into specific national and metro markets — precisely where small in-city footprints have the natural advantage.

Social license

Public acceptance has hardened into a structural variable. A Gallup poll in May 2026 found **71% of Americans oppose AI data centers in their area** — a higher rate of opposition than nuclear plants draw (Gallup). Data Center Watch tallied **$64B of US projects blocked or delayed** over the two years to March 2025, with opposition bipartisan (Data Center Dynamics).

The recurring flashpoints are water, with roughly two-thirds of US data centers sited in water-stressed areas; diesel and noise; and ratepayer cost-shifting. In Latin America, water is the sharpest fault line — Chile's environmental court partially revoked Google's Cerrillos permit over aquifer impacts in 2024, and Querétaro, home to roughly 79% of Mexican capacity, faces its worst drought in a century with no new water concessions being granted (Data Center Dynamics; Mexico Business).

The countervailing asset belongs almost exclusively to urban operators: waste-heat reuse. Microsoft and Fortum's partnership will heat roughly 250,000 people, and Nordic and German cities now actively invite data centers onto their district-heating networks (Fortum; Araner). An in-city facility feeding a district-heating loop converts waste heat into a tangible community benefit — the strongest social-license lever available, and one a remote campus cannot pull. In a regulatory environment where social acceptance increasingly gates approval, proximity to the city is not a liability. It is the case for being there.

The Competitive Landscape

The market for AI-compute infrastructure is not short of capital or operators. What it lacks is anyone optimizing for the place where inference demand is converging: the dense urban core. The same pattern surfaces in every segment. The largest players are racing in one direction (up-market, toward gigawatt campuses on cheap exurban land), an earlier wave of would-be challengers tried the opposite extreme (ultra-distributed micro-edge) and failed, and the productive middle (5-40MW metro nodes in cities) is being structurally vacated even as the demand signals point straight at it.

The colocation majors are leaving the small-urban tier

The listed colocation and wholesale operators are the segment's anchors, and they are consolidating their bets on scale. Equinix, the interconnection leader with roughly 270 IBX facilities across more than 30 countries, is targeting a capacity doubling by 2029, with its xScale hyperscale unit leasing capacity at 430MW. Digital Realty is bringing roughly 5GW online across 40 metros and has been explicit about "doubling down on its largest markets." NTT, CyrusOne (~674MW operational plus a 760MW Texas campus), Vantage (central to the Stargate Wisconsin buildout) and QTS are all hyperscale and wholesale builders working at the large end of the market (NTT DATA; ABI Research).

The trajectory is consistent across the cohort.

OPERATOR	DIRECTION OF TRAVEL	SCALE SIGNAL
Equinix	Doubling capacity by 2029; xScale hyperscale	~270 IBX; 430MW xScale leased
Digital Realty	"Doubling down" on largest markets	~5GW across 40 metros
CyrusOne	Hyperscale / wholesale campuses	~674MW + 760MW TX campus
Iron Mountain	Brownfield conversion at scale	Tripling toward 1.3GW

Two operators in this group are instructive for a different reason. CoreSite, owned by American Tower, holds the strongest urban-interconnection assets in the sector (the dense metro carrier hotels), which makes it the closest incumbent comparable on the urban side of the market. And Iron Mountain is the most relevant brownfield-retrofit analog: it built a data center business in part by converting existing real estate, and is tripling that footprint toward 1.3GW (Iron Mountain 8-K, SEC). Iron Mountain's trajectory matters because it shows that brownfield-to-data-center conversion is bankable at scale.

The takeaway is structural. The colocation majors validate both the demand and the brownfield-conversion playbook, but they are sprinting toward 100MW-to-gigawatt campuses. None of them optimizes for the sub-40MW, dense-metro footprint (DCD; Bisnow).

The edge graveyard: a cautionary track record

The most important lesson in the competitive history is also the cleanest. Operators that built meaningful metro-interconnect at real node sizes thrived. Operators that chased ultra-distributed micro-edge (the "tiny boxes everywhere" thesis) failed.

The winners scaled into AI capacity at credible node sizes and partnered with the demand side. EdgeConneX, owned by EQT, has grown roughly twentyfold since 2020 on a "hyperlocal to hyperscale" model and partnered with the neocloud Lambda on more than 30MW of high-density AI capacity in Chicago in 2025. DataBank runs 65-plus data centers across 25-plus metro markets, the largest US metro portfolio by count, and raised roughly $5B in eighteen months (EdgeIR; DataBank; TechCrunch). Cologix, backed by Stonepeak, operates 45-plus facilities across 13 North American edge markets and has raised $1.5B.

The graveyard is equally clear. Vapor IO raised more than $200M to place micro-edge in 50 to 500 metros and has since shifted hard to AI-RAN software, effectively abandoning the original density thesis. EdgePresence was absorbed by Ubiquity in 2023; EdgeMicro went dormant. The tower-base wave is the most pointed example: American Tower deployed roughly six tower-base edge sites between 2019 and 2021, slowed the buildout, and bought CoreSite for $10.1B instead (choosing scale over micro-edge). Crown Castle's chief executive put it plainly: "we're not interested in investing in data centers" (Crunchbase; Fierce Network; DCD).

The lesson generalizes. The viable unit is the metro node at 5-40MW with real interconnection (the DataBank, Cologix and EdgeConneX node size), not the sub-megawatt density that stalled. This is exactly the size at which BMETAL builds urban AI-compute nodes (modular and brownfield, deployed fast in the cities where intelligence is made) rather than the failed pattern of hundreds of tiny sites.

The neoclouds: the demand engine, not the rival in the urban tier

The GPU-cloud operators (the "neoclouds") need gigawatts faster than they can self-build, and they are already signing 20-40MW metro deals with edge operators. CoreWeave ran 850MW of active capacity in 2025, heading toward more than 1.7GW exiting 2026, against a $66.8B backlog. Crusoe raised a $1.375B Series E at a valuation above $10B and is building the 1.2GW Abilene Stargate campus. Nebius is targeting 800MW-to-1GW by the end of 2026 against 3.5GW contracted (Crusoe; Introl; AgentMarketCap).

Crucially, several neoclouds already buy edge capacity rather than build everything themselves. Lambda runs a 100MW Kansas City "AI Factory" and has partnered with EdgeConneX (the 30MW Chicago deployment) and with ECL in Houston. Voltage Park, which owns 35,000-plus GPUs and merged with Lightning AI in January 2026, leases shells. Together AI and Vast.ai run largely capacity-light, leased footprints. The structural read is that these are not facility competitors in the urban tier; they are the customers and channels that the urban tier serves.

Modular and prefab builders: adjacent, not overlapping

The modular and prefab cohort is the closest set of analogs, and it splits cleanly into two groups, neither of which targets dense-urban colocation.

OPERATOR	PRIMARY DIFFERENTIATOR	SITING FOCUS
Armada	Containerized → MW-scale "Leviathan" hardware	Military / remote / energy edge
ECL	Hydrogen-powered, off-grid	Exurban (1GW TerraSite-TX, Houston)
Verrus	Grid-interactive (compute + battery)	Power flexibility, not urban land
Soluna	Renewable-colocated	Remote / off-grid

Armada has raised roughly $469M (including a $230M Series B at a $2B valuation in May 2026, led by BlackRock) and builds containerized-to-megawatt-scale systems aimed at military, remote and energy-edge deployments rather than dense urban colocation (CNBC; Tech Startups). ECL builds hydrogen-powered, off-grid modular capacity (its 1GW TerraSite-TX sits in the Houston exurbs), so its differentiator is power and its siting is exurban. Verrus, a spinout from Alphabet's Sidewalk, competes on grid-interactive power flexibility (roughly 50MW of compute plus battery per building) rather than on urban land. Soluna sits at the opposite siting extreme: renewable-colocated and remote (Data Center Frontier; Latitude Media).

The structure here is the clearest signal of white space in the entire landscape. The modular wave divides into box manufacturers (Armada, oriented to defense and remote) and power-innovation operators (ECL, Verrus, Soluna, all non-urban). The operators that do own dense urban metro footprints (DataBank, Cologix) build conventional facilities rather than a standardized, modular, brownfield product. No well-funded player runs an urban, brownfield, modular multi-tenant model as a repeatable platform.

The tower-company template and the capital backdrop

The independent tower companies are a structural reference for how contracted, multi-tenant infrastructure scales. American Tower (~149,000 sites, and the owner of CoreSite), Crown Castle (~40,000 towers) and SBA (~40,000) together hold roughly 70% of privately owned US towers, run 5-7% annual escalators, and trade on contracted cash flow (Mordor Intelligence; celltowerleases.com). Whether a tower owner that already controls urban sites and a premier metro operator ever standardizes an urban modular product remains an open industry question; the earlier retreat from tower-base edge suggests the bar is high.

The capital backdrop is unusually deep. Data center M&A set records in 2025 (roughly $61-69B by various scopes, with US private equity alone at $45.7B, a five-year high). Aligned Data Centers sold for roughly $40B in October 2025, the largest data center deal on record and more than double the prior high (AirTrunk at ~$16.1B in 2024) (The Tech Capital; S&P Global; Data Center Knowledge). Capital, in short, is not the constraint. Where it flows next is the open question.

The white space

Mapping the field as industry structure, the gap sits at an intersection that no funded competitor currently occupies:

- **Sub-40MW, dense-urban siting.** The incumbents are fleeing upward into gigawatt campuses, structurally vacating the small-urban tier precisely as inference, latency and sovereignty pull demand back into cities.
- **The proven node size.** The edge graveyard demonstrates that "hundreds of tiny sites" fails; the winners run dozens of 5-40MW metro nodes. The viable position is at the winning size, not the failed one.
- **Brownfield and modular as the default unit.** Iron Mountain shows brownfield conversion is bankable; Armada, ECL and Verrus show modular is fundable. Nobody combines brownfield urban shells, standardized modular fit-out and multi-tenant operation as a single repeatable product.
- **Power solved at the site.** With grid interconnection queues stretching past 2035 in markets such as West London, the binding constraint is power. Behind-the-meter generation where the grid cannot deliver is part of the standard playbook for getting urban capacity online fast.

This is the physical layer of AI in the cities where intelligence is made: infrastructure-grade urban real estate and power, modular and brownfield compute nodes deployed fast at the node size the market has already proven,

with waste heat reused on site. The majors are building bigger and farther out. The micro-edge wave already tried smaller and lost. The dense-urban middle remains open.

Technology: Bare-Metal, Inference & Distributed Compute

The defining shift in AI infrastructure is no longer where the largest models are trained — it is where they are run. As inference moves from a research cost into the dominant production workload, the technology stack that serves it is being rebuilt around a different set of constraints: proximity to users, verifiable tenant isolation, and the ability to coordinate compute across many locations rather than one. This is the technical substance of urban AI-compute, and it is converging on five directions: bare-metal GPU provisioning, high-throughput inference serving, an increasingly diverse silicon layer beneath both, confidential computing for tenant trust, and latency-aware distributed inference across a network of urban nodes.

From one big cluster to many small ones

Nearly every mature piece of AI-infrastructure software shipped in 2025–26 — Kubernetes GPU operators, NVIDIA Dynamo, llm-d, Slurm, Run:ai — was engineered for a single assumption: one large, homogeneous cluster inside one building, with sub-microsecond NVLink inside a node and 400–800 Gb/s InfiniBand or RoCE across the room. That architecture is optimal for training, where a model campaign benefits from one tightly-coupled fabric.

Inference breaks the assumption. The economically interesting topology is not one warehouse-scale cluster but a network of smaller clusters — sub-5MW edge nodes through roughly 40MW colocation-class sites — distributed across a metro and, eventually, across metros. Across metro and wide-area links, the operating environment is fundamentally different: 1–10 ms or more of latency, variable bandwidth, and heterogeneous hardware. The industry's best tooling stops cleanly at the single-cluster boundary, which is precisely the boundary a distributed urban network is built to cross.

Bare-metal provisioning: a mature commodity layer

The provisioning layer — turning racked servers into a reschedulable pool of GPUs — is a solved problem, assembled from mature open source. Kubernetes with the NVIDIA GPU Operator handles orchestration; Slurm covers training workloads; OpenStack Ironic provisions bare metal. NVIDIA open-sourced Run:ai following its acquisition of the company, removing the last meaningful proprietary advantage at this layer and making sophisticated GPU scheduling broadly available (Tom's Hardware).

Bare-metal GPU rental itself has become a visible, cyclical commodity. H100 on-demand pricing fell roughly 64%, from about $8/hr in 2023 to a $1.70/hr one-year-contract low in October 2025, then rebounded around 40% to roughly $2.35/hr by March 2026, as more than 300 new providers flooded supply before early-2026 demand outran it (Introl; Silicon Data). The price of the silicon is converging across the market; what does not converge is location. For latency-sensitive inference, proximity to users — not headline $/GPU-hour — is where differentiation lives.

There is a structural pricing gradient worth noting. The spread between the cheapest neocloud and a hyperscaler runs roughly 5× for the same H100 silicon, with hyperscalers generally charging 2.5–4× neocloud rates (Spheron; aimultiple). North America accounts for roughly 88% of neocloud GPU-as-a-service revenue, concentrating demand in exactly the dense urban metros where AI is built (Introl).

Inference serving: competing on throughput, not kernels

The inference runtime — the software that actually executes a model on the GPU — is the subject of a well-funded open-source arms race. vLLM is the default; TensorRT-LLM delivers roughly 15–30% higher throughput on H100 at the cost of heavier operational complexity; SGLang excels at high-concurrency, prefix-heavy workloads (Spheron). The serving layer is not where the industry differentiates.

Token economics make this clear. Inference prices have fallen roughly 1,000× in three years — GPT-4-class serving went from about $20 per million tokens in 2022 to roughly $0.40 per million tokens in 2026, with open models running $0.05–0.88 per million tokens — driven by four compounding forces: hardware, serving software, model-architecture efficiency, and quantization (AI Pricing Guru). Independently sourced cost-to-serve sits around $0.95–1.10 per million tokens, which means generic token serving is margin-thin except at very high concurrency. The lever is throughput — more served demand per rack — not price. The throughput spread that custom silicon can buy is illustrated by Cerebras delivering roughly 3,000 tokens/second against Groq's 476 on gpt-oss-120B (IntuitionLabs).

This dynamic is enforced at the market level by aggregation. OpenRouter routes requests to the cheapest and fastest available provider, processing roughly 25 trillion tokens per week, with Chinese-origin open models making up an estimated 45–51% of all tokens routed and no single provider holding more than about 25% share (Dataconomy). On a generic open-model basis, providers compete in a real-time price-and-latency auction that arbitrages margin away. Differentiation in inference comes from latency guarantees, data residency, private and confidential deployment, and bundled capacity — never from the headline token price.

The silicon layer: inference-optimized accelerators and the multi-silicon era

The competitive frontier of inference is increasingly silicon diversity, not raw NVIDIA dominance. NVIDIA remains effectively unassailable in training, where a single tightly-coupled fabric and a roughly fifteen-year software moat compound into an enduring advantage. Inference is the flank. It is a more fragmented, more cost-sensitive, and more architecturally forgiving workload, and it is where alternative accelerators now compete directly on total cost of ownership rather than peak FLOPS.

Tenstorrent's Galaxy system, built on its Blackhole ASIC, is a useful exemplar of the class. A single 6U unit packs 32 Blackhole accelerators delivering roughly 23 PFLOPS of FP8 compute and 1 TB of GDDR6, and scales out over standard 800GbE Ethernet rather than proprietary NVLink or InfiniBand. The stack is open end to end — open RISC-V cores and an open-source software layer — with an entry price around $110k per unit and general availability expected mid-2026 (Tenstorrent; The Register; HPCwire). Significantly, a 6U Galaxy draws roughly 8–12 kW and is air-cooled.

The economic case is aggressive but must be read as vendor claims pending independent third-party benchmarks. Tenstorrent positions Galaxy at roughly a 5× inference TCO advantage versus NVIDIA's GB300-class systems — on the order of $6 versus $30 per million tokens — and cites roughly 350 tokens/second/user on a 671B-parameter reasoning model. These are vendor-supplied figures and have not yet been validated by neutral benchmarking houses (Spheron; Futurum; Moor Insights, 2026). The constraints are real. GDDR6, rather than HBM, caps the largest-context and highest-bandwidth workloads. More decisively, CUDA's roughly fifteen-year ecosystem is the genuine switching cost: porting production workloads off CUDA carries real engineering friction, and software maturity — not headline price — is the gating variable.

The urban relevance is direct. At roughly 8–12 kW per 6U, air-cooled, scaling over commodity Ethernet, this class of accelerator fits inside the power-and-cooling envelope of converted urban buildings. By contrast, NVIDIA's

GB300 NVL72, at roughly 120–140 kW per rack with mandatory liquid cooling, cannot drop into a Class-B retrofit without gutting HVAC and water systems. Silicon diversity therefore widens the set of urban shells that can host meaningful inference capacity, rather than confining it to the handful of sites that can support liquid-cooled, high-density racks.

This is not a speculative pattern. The single-vendor assumption has already broken at the frontier. Google operates large TPU fleets on its XLA/JAX stack, with expansion reported toward roughly one million TPUs and more than 1 GW of capacity, while AWS Trainium — programmed through its Neuron SDK — now runs in multi-hundred-thousand-chip clusters backed by roughly $8B of investment, both deployed alongside NVIDIA and used by frontier labs including Anthropic (Google, 2026; AWS, 2026). Non-NVIDIA silicon demonstrably wins at scale when software and economics align.

The nuance is where CUDA's moat actually binds. It holds the long tail of mid-market inference far more tightly than the frontier: hyperscalers and frontier labs employ compiler teams, while mid-market inference tenants do not. Near-term, the practical pattern is blended fleets — NVIDIA for CUDA-locked workloads, alternative accelerators for cost-sensitive inference — with managed-service abstraction carrying non-CUDA economics to mainstream tenants. The market signal is loud: Qualcomm has reportedly pursued Tenstorrent at up to roughly $10B (mid-2026), the kind of capital backing that turned AWS Trainium from a curiosity into frontier-grade infrastructure, and a marker that NVIDIA's inference flank is now contested ground (Reuters, 2026).

Confidential computing: hardware-rooted tenant trust

The most consequential trust development in the stack is hardware-native. H100 and Blackwell GPUs ship with a hardware trusted execution environment (TEE): an on-die root of trust, secure boot, attestation, and encrypted CPU-to-GPU transfer (NVIDIA). This lets a tenant cryptographically verify that even the infrastructure operator cannot read their model weights or data.

For an industry where workloads increasingly carry regulated, proprietary, or sovereignty-sensitive data, this answers the foundational question of whether to trust a third-party compute provider at all. The capability is hardware-native — low overhead in the single-digit to low-tens-of-percent range, depending on workload — yet remains under-marketed by most emerging providers. It is a feature that pairs with, rather than replaces, formal compliance posture such as SOC 2 and ISO 27001. As confidential computing reaches general availability across the current GPU generations, attestable isolation is moving from a niche enterprise request to a baseline expectation for sensitive inference.

Post-training: the second urban workload

Between frontier training and production inference sits the fastest-growing enterprise workload class: post-training — fine-tuning, distillation, and reinforcement-based refinement of open and small models on proprietary data. Unlike inference it is batch work with no latency constraint, and unlike frontier training it fits in racks, not campuses: production post-training runs routinely complete on tens to hundreds of accelerators rather than tens of thousands. What pulls it toward the city is not physics but control. The data it consumes is often the most sensitive an enterprise owns; the tuned weights it produces are crown-jewel intellectual property; and a growing share of it falls under the same sovereignty and localization rules traced earlier in this report. The natural delivery form is the dedicated private pod — sub-megawatt to a few megawatts of single-tenant, hardware-attested capacity inside an urban building, close to the data, the compliance boundary, and the team — sold on security and control rather than proximity alone. It pairs directly with the confidential-computing capabilities above, and the air-cooled accelerator class fits it naturally.

Disaggregated and cross-site inference

Inference serving is disaggregating: the phases of model execution — prefill and decode — are increasingly split across different accelerators, sized and priced for each phase, rather than run monolithically on one class of chip. It is the architecture the newest inference platforms are built around, and it rewards exactly the heterogeneous, multi-silicon fleets described above. The frontier of the stack is coordinating that across many sites. The leading single-cluster systems are excellent within their design envelope and fall apart outside it. NVIDIA Dynamo disaggregates prefill and decode and shuttles KV-cache state via a low-latency transfer library that assumes a microsecond-scale fabric; llm-d's KV-cache-aware router tracks state within a single Kubernetes cluster; LMCache assumes a shared fast transport (NVIDIA; llm-d; LMCache). Moving a KV cache across a metro or wide-area link, where latency is 1–10 ms or more, is catastrophic for these designs. Running them across sites simply does not work, and decentralized GPU networks that attempt fleet coordination tend to sacrifice reliability, isolation, and service-level objectives to do it.

A genuine cross-site inference fabric — presenting one logical endpoint while routing and placement happen invisibly across dozens of nodes — has to solve a distinct set of problems:

- **Global request routing** — directing each request to the node that minimizes user-perceived latency subject to capacity and SLOs: the nearest node that already has the model warm, not merely the nearest node.
- **Model placement** — deciding which models live on which nodes given each site's memory, demand geography, and cold-start cost, as a continuous optimization over time.
- **Cross-site state strategy** — generally avoiding KV migration across the wide area, instead using session affinity to route follow-on requests back to a warm node and replicating only hot prefixes such as system prompts and shared context.
- **Cold-start and model-loading management** — predictive pre-warming from demand signals, weight streaming from local cache, and scale-to-zero across sites.
- **SLO-aware admission and overflow** — spilling a saturated metro to a neighboring node or fallback capacity while honoring tenant service guarantees.

This is the direction the industry is moving as inference disperses toward the cities where it is consumed. The architectures that win will treat a distributed network of urban nodes as one latency-aware fabric — the capability that single-warehouse infrastructure, by design, cannot provide.

The technical shape of the opportunity

The urban AI-compute stack is, layer by layer, a story of consolidation at the bottom and open frontier at the top. Provisioning, serving runtimes, observability, and metering are mature and broadly available. Networking has shifted toward Ethernet — RoCEv2 and Spectrum-X close roughly 80–90% of InfiniBand's performance gap for inference-first workloads at lower cost, while exotic fabric is reserved for large training clusters (Spheron). What remains genuinely unsolved is the coordination layer: latency-aware inference across a distributed network of urban nodes, secured by hardware-rooted confidential computing. That is where the physical layer of AI is heading, and it is being built in the cities where intelligence is made.

Geography: Where Urban Compute Concentrates

AI compute is bifurcating. Training gravitates to remote gigawatt campuses where cheap power is abundant; inference is moving back toward cities, pulled by latency, cost, data gravity, and sovereignty. As one industry framing puts it, "training built the campuses; inference will choose the markets" (DCD). The question for the urban tier is therefore geographic: which metros combine deep AI demand with the power and regulatory conditions that let small, fast-deployed footprints win.

The answer is not uniform. Each candidate metro trades demand density against power availability, scarcity, electricity cost, and permitting friction. What follows is an analysis of where the urban AI-compute opportunity concentrates and the local trade-offs that define each market.

What separates an urban-compute metro from a hyperscale one

Hyperscalers optimize for cheap land and power at gigawatt scale, and are racing up-market into multi-hundred-megawatt and gigawatt campuses. That migration vacates the sub-40MW urban tier — the small footprints closest to where inference demand actually lives. The metros that matter for that tier privilege a different set of factors: proximity to AI demand, speed-to-power, and supply scarcity that confers pricing power. The leading commercial-real-estate trackers (CBRE, JLL, Cushman & Wakefield) consistently show the same pattern across these markets: record-low vacancy, lengthening interconnection queues, and demand concentrating in metros the incumbents are leaving behind.

The structural criteria that distinguish a strong urban-compute metro:

- **Inference and demand density** — proximity to the users, enterprises, and model developers that consume inference.
- **Power availability and interconnection speed** — the binding constraint in nearly every market.
- **Supply scarcity** — low vacancy signals pricing power and unmet demand pull.
- **Permitting and regulatory friendliness** — the determinant of deployment velocity.
- **Electricity cost** — important for operating margin, but secondary to availability.
- **Land cost and brownfield stock** — small footprints and lease models neutralize otherwise prohibitive metros.
- **Connectivity** — fiber, internet exchanges, and on-ramps, largely a hygiene factor in target metros.
- **Policy and incentive environment** — de-risks siting at the margin.

North America

North America's established cores are either saturated or hyperscaler-locked. Northern Virginia, the world's largest market, sits at roughly **0.5% vacancy** at a record **$196/kW** pricing benchmark (CBRE, H2 2025) — extreme scarcity, but already claimed at scale. Atlanta posted record-low vacancy near **1.6%**, but its growth is hyperscale-led (Georgia Recorder). Dallas and Phoenix remain open with cheaper power, but offer less of the scarcity arbitrage that defines the urban tier.

Two metros stand out for the urban tier — for opposite reasons.

New York City is the strongest demand-and-pricing story in North America. Colocation commands **$180–300+/kW/month** against a roughly **$163** national benchmark, and city officials have explicitly steered policy toward "smaller AI facilities, edge, and retrofits" (Construction Dive). The constraint is power: roughly 30 large connection requests landed between 2020 and 2025 against an aging Con Edison grid (The City), and both electricity cost and regulatory friction are among the worst in the country. The structural advantage for small footprints is precisely that megawatt-scale queues do not bind them — sub-5MW footprints can draw on existing building service and tap abundant brownfield stock in the outer boroughs and northern New Jersey, sidestepping the multi-hundred-megawatt interconnection backlog entirely.

The San Francisco Bay Area is the densest concentration of AI demand and talent on earth — home to the model, agent, and evaluation ecosystem, and the deepest pool of GPU and ML infrastructure expertise. Bay Area colocation runs at roughly **4% vacancy at the highest pricing in North America** (CBRE; Bisnow). The trade-off is power, on both cost and availability:

- **Cost:** PG&E all-in rates run roughly **$0.25–0.32/kWh**, with a commercial median up to about **$0.42** — three to four times the cheapest US markets (PG&E).
- **Availability:** PG&E interconnection runs multi-year, and Silicon Valley Power — the cheap municipal option — is effectively full, with roughly **100MW of finished capacity sitting idle awaiting grid energization to 2028–29** (Fortune).
- **Policy:** California's 2025 CEQA reforms (AB 130 / SB 131) exempted housing and advanced manufacturing but **explicitly not data centers** (Holland & Knight), keeping discretionary review in play; adaptive reuse within existing industrial entitlements is the path of least resistance. Bay Area air-quality rules (BAAQMD Tier-4) make low- and no-diesel designs a competitive advantage, and the region's mild marine climate makes liquid-cooled, economizer-friendly builds cheap to operate under California's Title 24.

San Francisco illustrates the broader urban pattern: world-class demand sits behind a binding power constraint. That constraint is why behind-the-meter generation — solid-oxide fuel cells (Bloom Energy, with precedent across Equinix's urban sites) paired with battery storage — is increasingly central to siting urban compute (Bloom Energy).

Western Europe

The FLAP cores — Frankfurt, London, Amsterdam, Paris — are maximally scarce but near-impenetrable for a newcomer. Frankfurt sits near **0.6% vacancy** with no land; Amsterdam has effectively no new grid connections available until roughly 2030; London faces long queues and high cost; and Ireland's regime imposes self-generation conditions on new connections (JLL, EMEA YE 2025; CRU). Scarcity in these markets is real but closed.

Madrid is the standout Western European urban-compute metro — the one major market combining genuine scarcity, strong demand, enterable economics, and friendly permitting:

- **Demand:** A top-10 EMEA hub with a **538MW pipeline**, a national IT-load forecast near **23.9% CAGR to 2030**, and submarine-cable gateway status linking Southern Europe to Latin America (Cushman & Wakefield).
- **Power:** Roughly **50% renewable generation** and a 2025 day-ahead average near **€65/MWh** — among the cheapest in Western Europe (OMIE). The constraint is a permit-to-energization gap: roughly **11.8GW of granted-but-not-operational** grid access, which favors operators targeting small loads in already-served corridors such as Alcobendas and Henares.

- **Regulation:** The most data-center-friendly major European metro, with a one-stop-shop permitting office and a strong national digital strategy (Mordor Intelligence).
- **Land and competition:** Cheaper than any FLAP core, with in-city scarcity pushing growth to nearby corridors. Equinix, Data4, Nabiax, and Merlin are established but undersupplied; hyperscale spend flows to Aragón, leaving the urban tier open. EU Energy Efficiency Directive compliance applies and shapes facility design.

The Nordics, by contrast, are excellent for training on cheap green power but weak on the urban inference density that the urban tier requires.

Latin America

Latin America presents the most extreme scarcity-versus-grid gaps in the world — the conditions under which behind-the-meter-powered footprints have the clearest structural advantage.

Querétaro and Mexico City form the most acute supply-demand imbalance in the analysis: Querétaro inventory grew **+450% year over year**, with the private sector self-building roughly **1,675MW because the national utility (CFE) cannot deliver** (Mexico Business; Tetakawi). The grid gap and CFE policy risk are real, but they are exactly the conditions that reward flexible, behind-the-meter power. Querétaro functions as a capacity hub; Mexico City, roughly 40 minutes away, supplies the inference demand — a market structure that naturally separates where power is available from where users are.

São Paulo is Latin America's deepest and cleanest-grid market:

- **Demand: 670MW operating plus 770MW queued** — about 42% of Brazil's total — supported by nearshoring and a large consumer digital base (IMAP; Scala).
- **Power:** A **94% renewable grid**, with Scala building a 560MW substation and federal tax incentives in play (Scala).
- **Regulation:** Sovereignty initiatives (Nuvem de Governo) drive in-country demand, though the grid is gated — the regulator (ANEEL) has approved 359MW that still needs three-to-five-year transmission builds.
- **Land and competition:** Cheaper than North America or Europe, with ample industrial and brownfield stock; Scala, Odata, and Ascenty are established but undersupplied against demand. The clean grid and comparatively lower energy-policy risk make São Paulo the most stable entry point in the region.

Santiago and Bogotá round out the regional picture — Santiago strong but smaller, with rising water-use pushback, and Bogotá among the most investment-friendly and least contested markets.

How the metros compare

Across the leading candidate metros, the structural criteria resolve into a consistent set of trade-offs.

METRO	DEMAND & TALENT	POWER	PERMITTING	STANDOUT CHARACTERISTIC
New York City	Strongest in North America; record colocation pricing	Aging grid, high cost, long queues	Among the most restrictive, but policy favors small footprints	World-class demand behind a binding power constraint

METRO	DEMAND & TALENT	POWER	PERMITTING	STANDOUT CHARACTERISTIC
San Francisco Bay Area	Densest AI demand and talent on earth	Highest US cost; municipal capacity effectively full	CEQA review in play; adaptive reuse the path of least resistance	Demand leadership offset by power scarcity
Madrid	Top-10 EMEA hub; submarine-cable gateway	Cheapest regional power; permit-to-energization gap	Most data-center-friendly major European metro	Cleanest balance of scarcity, demand, and speed in Western Europe
Querétaro / Mexico City	Strong, with capacity and demand split across two cities	Most acute grid gap in the analysis	Moderate; CFE policy risk	Extreme scarcity that rewards behind-the-meter power
São Paulo	Latin America's deepest market	94% renewable grid; transmission lead times	Friendly, with sovereignty-driven demand and incentives	Cleanest grid and lowest policy risk in its region

The pattern is consistent. New York and the Bay Area own the demand story but pay for it in power cost and grid friction. Madrid offers the cleanest balance of scarcity, demand, cheap regional power, and fast permitting in Western Europe. São Paulo pairs a near-fully-renewable grid with real incentives and the lowest policy risk in its region, while Querétaro and Mexico City represent the most extreme scarcity — the markets where flexible, behind-the-meter power converts a grid bottleneck into an opening.

Across every one of these metros the same logic holds: AI demand concentrates in cities, the grid cannot keep pace, and the incumbents are migrating to the gigawatt tier. The urban-compute opportunity sits in that gap — close to where intelligence is made, where speed-to-power and small footprints turn scarcity into advantage.

Risks to the Thesis

Every infrastructure thesis is a wager on a small set of variables holding. The case for urban AI-compute rests on three: that AI inference demand keeps compounding, that power can be brought to dense metros faster than the grid alone allows, and that policy keeps tilting the playing field toward small, distributed footprints. Each of these can break the other way. A credible analysis names the breaks plainly. What follows is the honest downside ledger for the urban-inference thesis — the macro forces that could weaken it, balanced against the structural reasons it may prove more resilient than the headlines suggest.

The demand question: is AI capex a bubble?

The single largest risk is exogenous and well-rehearsed: that the wave of AI infrastructure spending is running ahead of durable, monetizable demand. The numbers behind the boom are extraordinary — global data-center capex surged 57% in 2025, is projected to cross $1 trillion in 2026, and is forecast to grow at a 21% CAGR through 2029 (Dell'Oro Group, 2025). Hyperscaler capital outlays alone are running at $660–690B (TradeEdgePro / Chip Stock Investor). Skeptics, including high-profile investors, have flagged the circular financing patterns — chipmakers, clouds, and model labs cross-investing — and the aggressive depreciation assumptions underwriting much of this spend (IO Fund).

If AI capex normalizes, the consequences cascade through the colocation market that urban compute serves. The bear scenario is concrete: colocation pricing reverts from record levels, lease tenors shorten as tenants hesitate to commit, and the broader financing market sours on compute-backed debt. Record rents underwritten on multi-year pre-leases are exactly the kind of demand that looks fragile when the cycle turns.

Two things temper this risk without dismissing it. First, the demand drivers for inference specifically — as distinct from training — are structural rather than speculative: latency-sensitive applications, the cost of moving data, data gravity, and sovereignty requirements all push inference toward population centers. McKinsey projects inference growing at a 35% CAGR to exceed 50% of data-center demand by 2030 (McKinsey, The next big shifts in AI workloads), and Deloitte projects inference will account for two-thirds of AI compute as early as 2026 (Deloitte, TMT Predictions 2026). Inference demand tracks AI usage, not AI investment — and usage has proven far stickier than capex sentiment. Second, the supply backdrop is genuinely tight: Northern Virginia vacancy sits at 0.5% with record $196/kW pricing (CBRE, North America Data Center Trends H2 2025), and Bay Area colocation vacancy is around 4%. A demand cooling lands on a market with almost no slack, which cushions the price reversion the bear case fears.

The pace and cost of power build-out

Power, not land or capital, is the binding constraint on the entire sector — and the risk that it cannot be solved at urban sites on an investable timeline is real. The grid is the bottleneck: Dominion has acknowledged it cannot meet data-center power demand in Virginia, with connection queues stretching seven years (DCD); PJM's speed-to-power problem now exceeds eight years (RMI); and West London has been told there is "no capacity until 2035" (DCD). PJM's December 2025 capacity auction cleared at $333.44/MW-day while still coming up 6,625MW short (Utility Dive) — a market signaling acute scarcity.

Behind-the-meter generation answers this constraint — fuel cells and battery storage that bypass the interconnection queue — a market already approaching ~90GW of tracked capacity (Cleanview), validated at scale by deployments such as Oracle's 2.8GW and AEP's 1GW Bloom Energy commitments (Bloom Energy). But

this is a bet, not a certainty. If behind-the-meter economics disappoint — if turbine and fuel-cell lead times lengthen further amid the GE Vernova and Mitsubishi backlogs (Power Engineering), or if on-site generation proves costlier than expected — then the speed-to-power advantage that distinguishes urban sites narrows. Power cost itself is largely exogenous; it is a variable to design around, not one to bet on.

Regulatory tightening — and which way it cuts

Regulation is the most ambiguous risk in the ledger, because the same forces that constrain the sector can also advantage small urban footprints. The constraining edge is clear: the Netherlands has imposed hyperscale restrictions (DLA Piper), Ireland spent years under a connection moratorium (CRU), and the EU's Energy Efficiency Directive and Germany's EnEfG impose binding waste-heat and efficiency mandates (White & Case; EUR-Lex) that add real compliance overhead. Incentive regimes are being pulled back across US states, and large-load tariffs are proliferating (MultiState; Tax Foundation).

The countervailing point is that much of this regulation is structured around megawatt thresholds — the EED catches loads at ≥500kW and EnEfG colocation rules at ≥300kW, while hyperscale-specific bans and tariffs target the largest campuses. Threshold-based rules shape where small footprints have a structural advantage and large ones face friction. The genuine residual risk is social-license backlash that does not shrink with footprint: 71% of Americans oppose a data center nearby (Gallup; Data Center Watch via DCD), and noise, diesel, or water disputes can stall any urban project regardless of size. Heat-reuse offtake, closed-loop liquid cooling with near-zero water draw, and battery-over-diesel backup are the levers that convert this from a liability into a community asset — and in Europe, paid waste-heat lines (~2M SEK/MW at Stockholm Exergi) turn a mandate into a benefit (Stockholm Exergi).

Technology shifts: efficiency could shrink the compute pie

A subtler risk runs counter to the demand story: that model and serving efficiency improve fast enough to reduce the compute required per unit of intelligence delivered. The industry is moving quickly here — distributed and cross-site inference, smarter serving frameworks, and aggressive software optimization are all live directions. Falling GPU rental rates are the visible symptom: H100 rentals have already declined sharply from peak as supply and efficiency caught up (IntuitionLabs; Clouded Judgement). If efficiency gains outpace demand growth, the volume of physical compute the sector needs could grow more slowly than the headline forecasts imply.

The historical pattern, however, is the Jevons dynamic: cheaper compute has reliably expanded total consumption rather than contracting it. Token volumes bear this out — OpenRouter recorded on the order of 25 trillion tokens processed, with usage broadening across model providers (Dataconomy). Efficiency lowers the price of inference, which expands the set of economically viable applications, which raises aggregate demand. The risk is one of timing and distribution — which workloads run where — more than of the total addressable compute disappearing.

The historical edge-failure lesson

The most instructive risk is a cautionary tale the sector has already lived through. The first generation of "edge computing" — hundreds of tiny, sub-megawatt sites pushed as close to users as possible — largely failed commercially. Vapor IO, EdgeMicro, and early carrier and tower-company edge plays could not make the unit economics work at micro scale (Crunchbase; Fierce; DCD). The lesson is not that distributed compute is wrong; it

is that the granularity was wrong. Sub-megawatt density never reached the scale where interconnection, utilization, and operating leverage compound.

The urban-inference thesis is explicitly built against this failure mode. The viable footprint is not the sub-megawatt micro-site but the metro-scale node in the tens of megawatts — large enough to carry real interconnection density and operating economics, small enough to slip beneath hyperscale tariffs and zoning friction and to fit dense urban land. This is precisely the tier incumbents are vacating as they race up-market to gigawatt campuses. The edge failed at the wrong size; the white space is at the right one.

The shape of the downside

Read together, these risks are asymmetric in an informative way. The exogenous threats — an AI-capex correction, rate moves, behind-the-meter disappointment — are real and partly outside anyone's control. But the urban tier is structurally advantaged within each of them: a demand cooling lands on a near-zero-vacancy market; threshold-based regulation penalizes scale and spares small footprints; efficiency gains historically expand rather than contract aggregate demand; and the one prior attempt at distributed compute failed at a granularity the thesis deliberately avoids. As one industry observer put it, "training built the campuses; inference will choose the markets" (DCD). The risks to the thesis are genuine — but they are the risks of timing and execution within a structural shift, not of the shift itself.

Outlook to 2035

The forces mapped across this report do not point in competing directions. They point in one. Training, hungry for scale and cheap power, consolidates into rural and exurban campuses measured in the hundreds of megawatts to the gigawatt. Inference, governed by latency, cost, data gravity, and sovereignty law, is pulled the other way — back into the population centers where the demand actually sits. The split is structural, and it is durable. As the industry shorthand puts it: training built the campuses; inference will choose the markets (DCD).

The trajectory is set

The numbers behind that split are no longer speculative. Inference compute is on a trajectory of roughly 35% CAGR, growing from about 20.9 GW in 2025 to about 93.3 GW by 2030 — crossing over to become the majority of AI compute and an estimated 30–40% of total data-center demand by the end of the decade (McKinsey). Some read the curve as steeper still: Deloitte places inference at roughly two-thirds of AI compute as early as 2026 (Deloitte). Whichever pace holds, the destination is the same — a compute base whose center of gravity is the served user, not the training run.

The capital is moving to match. Data-center capital expenditure is projected to surpass $1 trillion in 2026 and to compound at roughly 21% through 2029 (Dell'Oro Group). Within that, global colocation alone runs near $77B in 2025 and is tracked toward roughly $256B by 2035. Capital at this scale does not reverse a structural trend; it accelerates the one already underway.

What fills the vacated tier

The opening this creates is on the supply side. Incumbents continue to vacate the sub-40MW, in-city tier as they chase gigawatt hyperscale — hyperscale operators hold roughly 48% of capacity today, a share projected to reach about 67% by 2031 (Synergy Research). The economics of the megacampus pull attention and balance sheets away from exactly the urban footprints that inference now needs most.

Several developments lower the barrier to filling that gap. Behind-the-meter generation is scaling into a credible answer where grids are full or interconnection queues are long, with roughly 90 GW now tracked (Cleanview). Air-cooled inference silicon lowers the retrofit bar, widening the universe of existing urban shells that can be brought into service without ground-up rebuilds. Closed-loop liquid cooling and waste-heat reuse are normalizing rather than remaining bespoke — a shift reinforced in Europe by the EnEfG and EED efficiency mandates, which turn heat recovery from a nicety into a compliance baseline. Each lowers the cost and shortens the timeline of putting compute inside the city.

The variable is pace, not direction

It is worth stating plainly what is and is not uncertain. Across bull, base, and bear scenarios, the direction is invariant: inference moves toward the user, urban supply stays scarce, and the economics of proximity continue to reward small, fast, in-city footprints over distant megacampuses. What differs between the scenarios is pace — how quickly power gets built, and how durable AI demand ultimately proves. A slower build-out stretches the timeline; it does not reverse the geography. The open variable is the clock, not the map.

The decade ahead

The synthesis is straightforward. The physical layer of AI is being rebuilt in cities. The compute that answers queries, serves agents, and runs regulated workloads will increasingly live near the people and the data it serves, even as the compute that trains the next model recedes to the rural campus. That bifurcation defines the build-out through 2035.

In a market where urban land is scarce, urban power is constrained, and the incumbents are retreating from the in-city tier precisely as demand for it accelerates, the structural advantage sits with operators able to secure land and power. That is the layer this report's publisher, BMETAL, works in: the physical layer of AI, in the cities where intelligence is made.

Sources & References

The analysis in this report draws on public market research, industry data, regulatory filings, and technology documentation current to 2025–2026. References are grouped by theme below.

Market sizing & AI-compute demand

- Dell'Oro Group — Data center capex surges 57% in 2025
- Dell'Oro Group — Data center capex to grow at 21% CAGR through 2029
- McKinsey — The cost of compute; The next big shifts in AI workloads
- Deloitte — TMT Predictions 2026
- Grand View Research — Global data center market
- Precedence Research — Global data center market
- Synergy Research — Hyperscale capacity share
- Expert Market Research — Colocation market
- MarketsandMarkets — Colocation and edge market
- Grand View Research — Edge Data Center Market Report
- Grand View Research — Modular Data Center Market Report
- Mordor Intelligence — Colocation and modular DC forecasts
- IMARC — Latin America DC market
- Arizton — Latin America colocation CAGR
- DCD — Training built the campuses; inference will choose the markets
- Tomasz Tunguz — Enterprise token-volume growth
- Dataconomy — OpenRouter token volume and model-provider share

Data sovereignty & geopatriation

- NetApp / Gartner — Data sovereignty and 75%+ geopatriation by 2030
- IDC — CIO repatriation intent (2025)

Capacity, vacancy & pricing

- CBRE — North America Data Center Trends H2 2025
- CBRE — Global Data Center Trends 2025
- JLL — 2026 Global Outlook; EMEA Data Centre Report YE 2025
- Cushman & Wakefield — EMEA Data Centre Update H2 2025
- Bisnow — Bay Area colocation; Equinix & Digital Realty pipelines

Power, grid & interconnection

- RMI — PJM's Speed-to-Power Problem
- DCD — Dominion Energy on Virginia data-center power demand
- Utility Dive — PJM December 2025 capacity auction
- PJM Inside Lines — December 2025 capacity auction results
- CRU (Ireland) — December 2025 data-center connection decision
- NL Times — Netherlands netcongestie / Amsterdam grid
- Mexico Business — Querétaro / CFE
- Energy Partnership Chile
- EIA — Electric Power Monthly
- UK DESNZ — Quarterly Energy Prices
- Eurostat / Statista — EU non-household electricity prices
- PG&E — Tariffs
- Fortune — Silicon Valley Power idle capacity
- OMIE / euenergy — Spain day-ahead power prices

Behind-the-meter generation & on-site power

- Cleanview — Behind-the-meter data center tracker
- Bloom Energy — Oracle, AEP and Equinix deals; solid-oxide fuel cells
- Power Engineering — Turbine backlogs
- Scale Microgrids — Solar acreage analysis
- Energy-Storage.News — BESS pricing

Density, cooling & waste-heat reuse

- Introl — Rack density
- Spheron — Rack density
- Dell'Oro Group — Liquid cooling market
- Uptime Institute — Global Data Center Survey 2025
- Nixon Peabody — Water legal and regulatory risk
- Stockholm Exergi — Heat recovery
- Fortum — Heat reuse
- Araner — Heat reuse and social license

Modular & brownfield build

- Bisnow — Brownfield retrofit
- Cabling Install — Telco central office conversion

- Bit Digital — North Carolina factory conversion

Regulation, zoning & policy

- MultiState — Large-load tariffs and incentive pullback
- Tax Foundation — Large-load tariffs
- Utility Dive — AEP, Dominion GS-5, Texas SB 6
- Holland & Knight — US data center zoning; CEQA AB130/SB131
- JLARC (Virginia) — US zoning review
- Smart Cities Dive — US zoning trends
- WTOP — Prince William Digital Gateway
- McGuireWoods — Prince William Digital Gateway
- EESI — Noise ordinances
- GPB — Chandler, AZ noise ordinance
- EUR-Lex — Energy Efficiency Directive and Delegated Regulation 2024/1364
- European Commission — EU energy efficiency
- White & Case — German EnEfG ERF mandates and LatAm regimes
- DLA Piper — Netherlands and Frankfurt rules
- CMS — Netherlands hyperscale rules
- Pinsent Masons — Ireland conditions
- Morgan Lewis — Singapore DC-CFA2
- IMDA — Singapore DC-CFA2
- Bird & Bird — Frankfurt Masterplan
- EY — LatAm data regimes

Social license & public opposition

- Gallup — AI data center opposition (2026)
- Data Center Dynamics — Data Center Watch; Cerrillos; Querétaro

Competitive landscape

- DCD — Q4/Q3 2025 colocation results; 2025 M&A; Crown Castle CEO
- NTT DATA — Hyperscale agreements
- ABI Research — CyrusOne 674MW
- Iron Mountain 8-K (SEC) — 452MW to 1.3GW
- Cushman & Wakefield — EMEA Data Centre Update H2 2025 (FLAP-D vacancy)
- EdgeIR — Cologix $1.5B
- DataBank — Portfolio

- TechCrunch — DataBank capital raise
- Crunchbase — Vapor IO
- Fierce Network — American Tower edge
- Crusoe — Capacity and funding
- AgentMarketCap — Neocloud capacity
- CNBC — Armada $230M/$2B
- Tech Startups — Armada
- Data Center Frontier — ECL; Verrus
- Latitude Media — Verrus
- Mordor Intelligence — Towerco model and escalators
- celltowerleases.com — Towerco escalators
- The Tech Capital — Aligned $40B; PE $45.7B
- S&P Global — 2026 outlook
- Data Center Knowledge — DC M&A and outlook

Compute economics & GPU markets

- Introl — GPU cloud price collapse and rebound; neocloud unit economics
- Silicon Data — H100 rental price over time
- Spheron — GPU cloud pricing comparison 2026
- aimultiple — GPU price index
- AI Pricing Guru — Token pricing trends
- IntuitionLabs — Cerebras vs SambaNova vs Groq; custom-silicon context
- Clouded Judgement — Neocloud revenue/MW; H100 rental deflation
- IO Fund — Circular financing in AI infrastructure
- TradeEdgePro / Chip Stock Investor — Hyperscaler capex

Inference & distribution technology

- Tom's Hardware — NVIDIA finalizes Run:ai acquisition and open-sources it
- Spheron — vLLM vs TensorRT-LLM vs SGLang benchmarks
- Spheron — GPU networking: InfiniBand, RoCE, Spectrum-X guide
- NVIDIA Developer — Confidential Computing general access on H100
- NVIDIA Developer — Introducing NVIDIA Dynamo (distributed inference)
- llm-d — Cross-cluster inference router
- LMCache — KV-cache transport technical report

AI silicon & accelerators

- Tenstorrent — Galaxy / Blackhole systems and specifications
- The Register — Tenstorrent Galaxy specifications, pricing and availability
- HPCwire — Tenstorrent Galaxy production volumes and deployments
- Spheron — Tenstorrent vs NVIDIA inference TCO comparison
- Futurum Group — Tenstorrent Galaxy and inference-silicon analysis
- Moor Insights & Strategy — Tenstorrent inference economics (2026)
- Reuters — Qualcomm–Tenstorrent acquisition interest
- Google Cloud — TPU scale and capacity expansion (2026)
- AWS — Trainium and Project Rainier (2026)

Geography & urban markets

- Construction Dive — NYC edge/retrofit policy
- The City — NYC grid and power
- Mordor Intelligence — Spain DC market and power
- Tetakawi — Querétaro insights
- IMAP — São Paulo market
- Scala Data Centers — São Paulo grid and substation
- Georgia Recorder — Atlanta data center growth