BEARINGLABS

Orbital Compute Frontier

Executive Summary

The ground remains the commodity-compute benchmark through at least 2035 and remains the base case for 2040.

Executive summary: the ground still wins the commodity-compute race

The ground remains the commodity-compute benchmark through at least 2035 and remains the base case for 2040.

The public case for putting AI compute in orbit has become more serious because the terrestrial data-center system is under visible stress: power queues are long, transformer lead times are measured in years, water and generator emissions have become political issues, and mature markets such as Loudoun County are tightening data-center approvals. But the decisive comparison is not orbit versus today's worst grid bottleneck. It is orbit versus the terrestrial frontier that is already adapting: secondary power-rich markets, utility-backed transmission, large-load tariffs, behind-the-meter generation, batteries and flexible load, direct-to-chip cooling, dry/hybrid heat rejection, reclaimed water, modular construction, and rapid silicon refresh.

The verdict is clear: for general-purpose AI training and inference serving terrestrial users, terrestrial data centers remain the best answer on power, schedule, connectivity, and cost through at least 2035; they are still the base case for 2040. Space-based compute has real niches—orbital edge processing, defense and sovereignty workloads, space-to-space services, grid-constrained premium compute, and brand or strategic optionality—but it does not yet clear the cost-per-watt hurdle for bulk AI compute.

Five findings drive that conclusion:

1. The terrestrial bottleneck is real, but it is not static. Lawrence Berkeley National Laboratory reports U.S. data centers consumed about 176 TWh in 2023 and could consume 325–580 TWh in 2028, or 6.7%–12% of U.S. electricity consumption (LBNL 2024 United States Data Center Energy Usage Report). LBNL's Queued Up work shows nearly 2,600 GW of generation and storage capacity seeking grid connection at year-end 2023 (LBNL / OSTI). Those constraints explain the orbital pitch. They do not prove it.
2. The modeled terrestrial parity target is roughly $12.6 per present-value usable IT-watt-year. A 1 GW facility-load terrestrial frontier campus, modeled as roughly 620 MW of usable IT load, produces a present-value cost intensity of about $12.6 per usable IT-watt-year under the stated assumptions. A near-term orbital case modeled at roughly $80.3 per usable IT-watt-year is about 6.4× the terrestrial baseline; an aggressive 2035 orbital steelman at $40.0 is about 3.2×; even a radical 2040 low-cost case at $23.7 remains about 1.9×. The parity rule is simple: an orbital stack has to deliver the same usable compute for less than roughly $46.5 billion PV cost in this 620 MW usable-IT reference case, including launch, spacecraft, power, radiators, communications, spares, refresh, insurance, deorbit, and ground segment. The discount-rate anchor is a 10-year Treasury around 4.45% from the U.S.

Treasury daily yield curve, built into an illustrative 9.2% infrastructure WACC (U.S. Treasury daily rates CSV).

3. Launch cost matters, but it is not the whole problem. A NASA paper cites a Starship-style aspiration around $10/kg to orbit, a figure that should be treated as aspirational rather than bankable (NASA NTRS, "Take or Make in Space"). Once launch price falls below roughly the low-hundreds of dollars per kilogram, non-launch costs dominate: flight-qualified compute, solar power, power conversion, radiators, spacecraft bus, optical communications, station-keeping, insurance, spares, ground systems, and hardware refresh. The cheapest credible orbital story therefore has to be a mass-manufactured module story: repeatable compute, power, thermal, communications, and bus units produced at satellite-factory scale—not merely a cheaper-rocket story.

4. Orbit's cleanest advantage is water, but its hardest problem is heat. A 1 GW terrestrial IT load using a hybrid evaporative profile of 0.40–1.20 liters/kWh-IT would consume roughly 3.5–10.5 billion liters per year, or about 2.5–7.6 million gallons per day. Orbit avoids that local water burden. But in orbit, essentially all IT power still becomes heat, and heat has to be radiated. A first-order Stefan-Boltzmann sizing for 1 GW of IT heat gives roughly 2.6 million m^2 of radiator at 300 K, 2.0 million m^2 at 320 K, 1.4 million m^2 at 350 K, or 0.8 million m^2 at 400 K, before mass, deployment, degradation, redundancy, orientation, solar absorption, albedo, and planetshine penalties. NASA describes spacecraft radiators as dedicated high-emissivity surfaces for rejecting heat (NASA Small Spacecraft Thermal Control); NASA technical material uses the Stefan-Boltzmann constant in the same heat-transfer form used here (NASA NTRS radiator source).

5. Space is regulatory and political arbitrage, not escape. Terrestrial data centers face zoning, stormwater, wetlands, air permits, utility service, transmission, and community politics. Orbital compute shifts the forum to the Federal Communications Commission, International Telecommunication Union coordination, FAA launch licensing, FCC orbital-debris rules, NOAA remote-sensing licensing if imaging is involved, and export controls. FCC's Part 25 satellite process (FCC Part 25), FAA Part 450 launch licensing (FAA launch/reentry licensing), FCC's five-year low-Earth-orbit deorbit rule (FCC deorbit rule), and FCC/ITU coordination (FCC international satellite coordination) are not trivial substitutes for county hearings.

Single biggest risk to this recommendation: a SpaceX/Starship-enabled step-change in launch cadence, coupled with mass-produced orbital power, thermal, and compute modules, could prove usable compute watts at terrestrial-like refresh economics before terrestrial grid and cooling innovation continues compounding. SpaceX's public filing for a proposed orbital data-center system is therefore the watch item that matters most. FCC public notice materials state SpaceX filed for an NGSO system of up to one million satellites between 500 km and 2,000 km, using high-bandwidth optical inter-satellite links (FCC SpaceX public notice). That is a serious signal. It is not deployed economic capacity.

1. Why the question exists now

The orbital-compute question exists because the terrestrial compute buildout has become an energy-infrastructure problem. The old assumption—that data centers could be absorbed as another commercial load class—has broken down at AI scale.

LBNL's 2024 data-center energy report estimates U.S. data-center electricity consumption rose to 176 TWh in 2023 and could reach 325–580 TWh in 2028 (LBNL report). That is not a marginal facilities-management problem; it is a generation, transmission, equipment, rate-design, and land-use problem. LBNL's 2024 interconnection-queue work separately reports nearly 2,600 GW of generation and storage capacity seeking grid connection at the end of 2023, with most projects that enter the queue ultimately withdrawn (LBNL / OSTI).

Equipment lead times make the constraint tangible. CISA's National Infrastructure Advisory Council report cites 80–210 week lead times for large transformers (CISA / NIAC transformer report). The Department of Energy's large-power-transformer resilience report separately frames extended replacement lead times as a grid-resilience issue, with maximum lead times reaching as much as 60 months in some contexts (DOE Large Power Transformer Resilience Report). Those are not contradictory numbers; they describe a range of transformer classes, procurement situations, and maximum-versus-typical planning cases. The implication is the same: transformer procurement can set the schedule.

Local politics are tightening as well. Loudoun County's data-center standards page states that the county moved to designate data centers as conditional or Special Exception uses in areas where they had previously been allowed by right (Loudoun County Data Center Standards & Locations). Virginia's Joint Legislative Audit and Review Commission has also identified data-center load as a driver of generation and transmission buildout difficulty, customer-cost exposure, and backup-generator emissions (Virginia JLARC presentation).

This is the strongest version of the orbital argument: terrestrial compute needs land, water, generators, substations, transmission, transformers, local permits, local political consent, and sometimes years of queue exposure. Orbit offers continuous solar flux, no local cooling-water withdrawal, no county zoning hearing for the compute platform, and a new venue for capacity growth.

That strongest version still has to answer the only question that matters: can orbit deliver usable compute watts, at the relevant hardware vintage, below the cost and schedule of the improving ground-side frontier?

2. The space-based compute landscape: status, not hype

The market is best understood in three waves.

Wave 1: orbital edge processing is real

In-orbit computing is not speculative. NASA describes HPE sending a high-performance computer to the International Space Station in 2017 (NASA Spinoff, Spaceborne Computer). NASA also describes SpaceCube onboard processors and the SCENIC system launched to the ISS in March 2023 (NASA SpaceCube). This matters for Earth-observation filtering, onboard autonomy, defense and intelligence workloads, and space-to-space services.

It does not prove that orbital hyperscale AI training is economic. The difference between onboard edge processing and a 1 GW AI data-center substitute is not incremental; it is orders of magnitude.

Wave 2: limited orbital GPU or hosted-compute service is emerging, but not proven

Axiom Space is a credible LEO infrastructure actor. NASA has described Axiom's commercial station pathway, including a Payload, Power, and Thermal Module that may support a free-flying Axiom Station as early as 2028, followed by habitation, airlock, and research/manufacturing modules (NASA Axiom station update). NASA also documents Axiom private astronaut missions to the ISS (NASA Ax-2 release). That makes Axiom relevant to hosted payloads and premium orbital services, not yet to commodity AI compute.

Kepler Communications is relevant to orbital communications infrastructure rather than compute capacity. The FCC granted U.S. market access for Kepler's proposed 140-satellite LEO fixed-satellite-service system (FCC Kepler order). Backhaul and space networking matter, but a communications network is not a data center.

Lumen Orbit surfaced in FCC experimental-license materials for Lumen-1 (FCC attachment), but the public record reviewed here does not verify meaningful deployed GPU capacity, commercial pricing, SLAs, routine refresh, or a scaled fleet.

Wave 3: GW-scale orbital data-center concepts are filing-stage

The strongest scale signals are regulatory filings, not operating systems.

Starcloud's FCC public notice states the company sought authority for up to 88,000 satellites in sun-synchronous orbit between 600 km and 850 km as a distributed data center in space (FCC Starcloud public notice). That is material because it defines an architecture at a scale that could, in theory, matter to hyperscale computing. It is not evidence of deployed compute capacity.

SpaceX's orbital data-center filing is the more consequential watch item because SpaceX has unmatched launch and satellite-manufacturing experience. FCC materials state SpaceX filed for an NGSO orbital data-center system of up to one million satellites at 500–2,000 km, with high-bandwidth optical inter-satellite links (FCC SpaceX public notice; FCC Space Bureau notice). That is the actor most capable of invalidating a conservative ground-ahead forecast. But a filing is not a factory, a power system, a radiator architecture, a spectrum grant, an insurance market, a deorbit solution, or a refresh cadence.

For Google Project Suncatcher, Sophia Space, Cowboy Space, and Crusoe as orbital-compute actors, the reviewed source base did not verify flown orbital compute hardware, capacity, orbit, funding, or regulatory posture sufficient to affect the 2030–2040 economic recommendation. Crusoe remains important in terrestrial AI infrastructure, but public orbital evidence was not established in this analysis.

3. The terrestrial innovation frontier is the real comparison set

Space-based compute should not be compared with a stalled substation request in a saturated market. It should be compared with what the terrestrial system is already doing to respond.

Grid-side adaptation

The terrestrial frontier has at least six active levers.

Secondary-market siting. Hyperscale development is already moving toward markets with cheaper power, available land, utility willingness, and large tax bases. Reference-class evidence from large data-center utility proceedings shows the pattern: constrained hubs push sponsors toward staged capacity, secondary power-rich markets, direct infrastructure cost assignment, readiness scoring, and long-term service commitments rather than binary "build here or go to space" decisions. The Meta Richland Parish record, including Entergy Louisiana's Mount Olive–Sarepta transmission materials, illustrates the timing mismatch between 18–24 month data-center builds and three-to-five-year generation or seven-to-ten-year

transmission development, and identifies secondary markets with reliable cheap power as an explicit response.

Large-load tariffs and cost allocation. AEP Ohio's data-center tariff page states that the Public Utilities Commission of Ohio adopted a data-center tariff settlement in July 2025 (AEP Ohio Data Center Tariff). Dominion Energy Virginia's GS-4 schedule illustrates the conventional large-load service framework, including applicability once measured demand reaches or exceeds 500 kW in at least three billing months (Dominion Schedule GS-4). The direction is clear: large compute loads will increasingly pay for commitment, exit risk, and system upgrades.

Behind-the-meter and bridge power. Gas turbines, reciprocating engines, batteries, backup fleets, staged utility service, and dedicated substations are not elegant, but they are deployable. They are also politically exposed, especially where air permits and emissions become visible.

Grid-enhancing technologies. DOE states that dynamic line ratings can allow lines to deliver 50% more energy than labeled limits under favorable conditions (DOE grid-enhancing technologies). DOE also identifies dynamic line ratings, power-flow control devices, and analytical tools as grid-enhancing technologies (DOE GET R&D). These tools will not create a 1 GW interconnection by magic. They can, however, improve the terrestrial frontier at the margin and defer or optimize upgrades.

Virtual power plants and flexible load. DOE describes virtual power plants as potentially providing 80–160 GW of flexible capacity by 2030, addressing 10%–20% of peak load and saving about $10 billion/year in grid costs (DOE VPP projects). AI training is not infinitely flexible, but some workload scheduling, UPS/BESS dispatch, and non-critical curtailment can reduce grid stress.

Early transformer and equipment reservation. The transformer problem is solvable only by early procurement, long-horizon utility planning, and sponsor credit. It is a severe terrestrial disadvantage versus a purely conceptual orbital facility—but orbital compute still needs power electronics, flight hardware, ground stations, and launch cadence. The supply chain changes; it does not disappear.

Data-center-side adaptation

Cooling and facility design are also moving.

DOE/FEMP defines water usage effectiveness as annual site water use divided by IT equipment energy use in liters per kWh (DOE FEMP cooling-water efficiency). LBNL describes liquid cooling and notes cold-plate heat capture around 50%–60% in one cited setting, as well as immersion cooling with dielectric fluids (LBNL Liquid Cooling). DOE's best-practices guide

includes heat reuse as a design lever (DOE Best Practices Guide for Data Center Design). DOE/FEMP also describes thermosyphon hybrid cooling as a water- and cost-efficient hybrid approach (DOE FEMP thermosyphon hybrid cooling).

The terrestrial frontier is therefore not "evaporative towers forever." It is direct-to-chip, immersion, dry/hybrid cooling, reclaimed water, better controls, tighter PUE/WUE, and heat reuse where local offtake exists.

4. Economics head-to-head

The economic model used here is intentionally transparent and conservative toward orbit. The figures below are scenario-model outputs, not audited bids. Where no public market quote exists for a GW-scale orbital compute stack, the model states the assumption being tested and uses public anchors for order of magnitude.

Terrestrial baseline

The reference case is a 1 GW facility-load AI campus with roughly 620 MW usable IT load, 85% utilization, a three-year build, and 12 operating years. The model assumes approximately $24.8 billion of initial terrestrial capex: about $18.6 billion for accelerators, servers, storage, networking, racks, and IT-side electrical integration, modeled at roughly $30/W usable IT, and about $6.2 billion for land, building, power, cooling, and interconnection at roughly $10/W usable IT. Annual operating cost is modeled around $1.6 billion, with major IT refreshes in years 7 and 11.

Those assumptions are triangulated, not claimed as a single audited bill of materials. Public anchors include DOE's Frontier supercomputer contract described as more than $600 million for a system exceeding 1.5 exaflops (DOE Frontier announcement), AWS's announced $11 billion Indiana data-center campus (Indiana Economic Development Corporation), Google's $15 billion Missouri infrastructure investment announcement (Missouri Governor release), and DOE's announcement involving 10 GW of generation and 10 GW of data-center development with SoftBank and AEP Ohio (DOE announcement). These sources are not perfect comparables; they establish order of magnitude and public-market context.

The modeled terrestrial present-value cost is $46.5 billion, or roughly $12.6 per present-value usable IT-watt-year.

Orbital 2030 near-term case

The 2030 orbital case assumes near-term launch economics, high spacecraft manufacturing cost, conservative mass for the full orbital compute module—compute hardware, power generation and conversion, radiators, bus, communications, station-keeping, shielding, redundancy, deorbit hardware, and ground segment—plus spares, insurance, and refresh penalties. Its modeled PV cost is $296.1 billion, or roughly $80.3 per usable IT-watt-year—about 6.4× terrestrial.

Orbital 2035 steelman

The 2035 steelman uses aggressive reusable-launch assumptions and spacecraft mass-production learning. Its modeled PV cost is $147.5 billion, or roughly $40.0 per usable IT-watt-year—about 3.2× terrestrial. This is the most important case because it gives the orbital bull thesis real credit. It still does not win.

Orbital 2040 radical case

The 2040 radical low-cost case assumes very low launch cost, lighter mass per usable IT watt, lower spacecraft dollars per kilogram, and lower operating/insurance cost. Its modeled PV cost is $87.2 billion, or roughly $23.7 per usable IT-watt-year—about 1.9× terrestrial.

That means the radical 2040 case still needs an additional roughly 47% PV-cost reduction to match the terrestrial baseline. The frontier can move on both sides, but terrestrial also benefits from chip efficiency, power-market adaptation, cooling gains, modular construction, and global siting arbitrage.

5. The launch-cost dependency

Launch cost is the most visible variable because it is easy to quote as dollars per kilogram. It is not the decisive variable by itself.

NASA historical launch-cost material documents how expensive prior systems were, including very high Shuttle-era cost per kilogram (NASA NTRS launch-cost source). The Starship bull case relies on the idea that reusable heavy lift collapses launch cost and makes mass in orbit cheap enough for industrial-scale infrastructure. The strongest version of that view deserves attention.

But orbital compute is not a barrel of inert mass. It is flight-qualified compute, radiation tolerance, power generation, power distribution, thermal rejection, structures, comms, propulsion, collision avoidance, redundancy, spares, deorbit, insurance, and ground operations. Launch reductions have diminishing returns once launch is no longer the largest line item. In the 2035 steelman and 2040 radical cases, cheaper launch helps materially but does not close the gap because the non-launch stack remains too heavy and too expensive.

A useful decision tree follows:

LAUNCH ENVIRONMENT	WHAT CHANGES	VERDICT FOR COMMODITY AI COMPUTE
Falcon-class / near-term ride-share economics	Launch remains a major cost and schedule item	Not competitive by 2030
Reusable heavy-lift with low-hundreds $/kg	Launch improves; spacecraft, power, thermal, and refresh dominate	Still likely above terrestrial in 2035
Single-digit to tens $/kg plus high launch cadence	Launch ceases to be the bottle-neck	Only matters if mass-produced orbital compute, power, and thermal modules also collapse in cost

This is why SpaceX is the key risk to the recommendation: only a company with credible reusable launch, satellite factory experience, optical networking, and regulatory ambition can plausibly attack several variables at once. Even then, the public evidence is still filing-stage for orbital data centers.

6. Thermal and water

Orbit's water story is compelling. It avoids municipal supply commitments, groundwater and surface-water withdrawals, cooling-tower evaporation, blowdown discharge, local sewer acceptance, and many water-capacity approvals. In water-stressed basins, those avoided burdens are not trivial.

DOE/FEMP explains that cooling-tower makeup water replaces evaporation, blowdown, and drift, with blowdown required because dissolved solids remain as water evaporates (DOE FEMP Cooling Tower Management). A 1 GW continuous IT load using 0.40–1.20 L/kWh-IT consumes roughly 3.5–10.5 billion liters per year. That is politically material.

But water avoidance does not equal thermal free lunch. In space, there is no convective atmosphere and no cooling tower. Heat rejection is radiation. The idealized radiator area is:

$A = Q / (\varepsilon \sigma T^4)$

Using $\varepsilon = 0.85$ and $\sigma = 5.67 \times 10^{-8}$ W/m²–K⁴, 1 GW of IT heat requires about:

RADIATOR TEMPERATURE	IDEAL RADIATOR AREA FOR 1 GW IT HEAT
300 K	~2.6 million m²
320 K	~2.0 million m²
350 K	~1.4 million m²
400 K	~0.8 million m²

Higher radiator temperature reduces area by the fourth power, but it also raises equipment, coolant, reliability, material, and operating constraints. Real design must also handle solar absorption, albedo, planetshine, orientation, shadowing, degradation, micrometeoroids, pumps, manifolds, redundancy, and deployment failure. NASA identifies solar absorptivity and infrared emissivity as controlling optical properties for spacecraft thermal systems (NASA SOA Thermal Systems PDF). NASA's solar science material places total solar irradiance near 1,361 W/m² outside the atmosphere and Earth's globally averaged solar input around 340 W/m² (NASA Goddard solar irradiance). The raw solar advantage is real; the usable compute advantage is not automatic.

7. Regulation and politics: venue shift, not escape

A terrestrial AI data center generally works through local zoning, site plan, building permits, construction stormwater coverage, wetlands permits if waters are affected, air permits for backup or behind-the-meter generation, utility service, substation/transmission work, and certificate of occupancy. EPA states construction disturbing one acre or more typically requires Clean Water Act construction-stormwater permit coverage (EPA NPDES construction stormwater). EPA states Clean Water Act Section 404 regulates discharge of dredged or fill material into waters of the United States, including wetlands (EPA CWA 404). EPA also identifies stationary combustion turbines and engines—common data-center power sources—as subject to Clean Air Act standards (EPA Clean Air Act resources for data centers).

A clean, by-right terrestrial site can be permit-ready in roughly 6–12 months. A site requiring special-use approval, rezoning, substation upgrades, or contested site-plan review can move into 18–36 months. A project with individual CWA 404 permitting, major-source air permitting,

major transmission, or nuclear-backed generation can exceed 36 months, and new nuclear capacity is a multi-year licensing and construction path through the Nuclear Regulatory Commission (NRC licensing process).

Orbital compute has a different stack. It needs FCC Part 25 licensing or market access (FCC Part 25), FCC/ITU spectrum coordination (FCC coordination), FAA launch/reentry licensing under Part 450 (FAA licensing; 14 CFR Part 450), orbital-debris compliance including the five-year LEO deorbit rule (FCC deorbit rule), NOAA licensing if private remote sensing is involved (NOAA CRSRA licensing), and export-control classification under ITAR/EAR for controlled spacecraft, technical data, foreign persons, suppliers, or launch arrangements (State Department ITAR rule; Commerce BIS EAR rule).

A small demo payload using conventional spectrum and an already-licensed launch vehicle can be a 9–18 month regulatory exercise. A commercial LEO compute constellation is more plausibly 24–60 months. A novel high-power, large-constellation architecture with contested spectrum and debris concerns can stretch longer.

The political risk also shifts. Loudoun and Virginia show the terrestrial pattern: local approvals, transmission, ratepayer costs, generator emissions, and visual impacts. The Dalles water-use controversy illustrates how water transparency can become a durable issue; Stanford's And the West reported that Google's data center accounted for a quarter of the city's water use in 2021 after a disclosure dispute (Stanford And the West).

Orbital compute faces launch-site politics, spectrum incumbents, debris and collision risk, dark-sky and astronomy concerns, export-control scrutiny, and national-security review. FAA materials for SpaceX Starship/Super Heavy at Boca Chica show launch-site environmental review and public engagement as live process items (FAA SpaceX Starship stakeholder engagement). NOIRLab describes work with the International Astronomical Union-linked center to mitigate negative interference from large satellite constellations (NOIRLab dark-sky work).

The political conclusion is not that one path is clean and the other dirty. It is that terrestrial risk is local and utility-centered; orbital risk is federal, international, launch-site, and commons-centered.

8. First-of-kind risk

The orbital bull case is a first-of-kind infrastructure case. That matters because FOAK projects are systematically overconfident when they treat component estimates as if system integration were already solved.

Reference-class evidence is unforgiving. A U.S. GAO report on ITER found that DOE's cost and schedule estimates could not be used to set a performance baseline in part because they were linked to an unreliable international schedule and an uncertain funding plan. A later DOE budget justification for ITER stated that it had not been possible to confidently baseline the project due to international schedule delays, design and scope changes, funding constraints, regulatory requirements, risk mitigations, and management issues, and that DOE anticipated a substantial increase in total project cost once the project was ready to be baselined.

The lesson applies directly to orbital compute. A bottom-up model that prices launch, panels, compute, radiators, and comms as separable line items will understate risk unless it also prices system integration: radiation and single-event effects at scale, on-orbit servicing, sustained launch cadence, fleet collision management, debris disposal, insurance, satellite manufacturing yield, optical backhaul, hardware refresh, regulatory conditions, and customer workload migration.

The unproven orbital variables are not peripheral:

- GW-class deployable radiators and thermal control;
- mass per usable compute watt after power, shielding, bus, and redundancy;
- radiation-tolerant accelerator performance and failure rates;
- high-cadence launch without long standdowns;
- satellite factory throughput for initial deployment and refresh;
- spectrum and optical backhaul at compute-scale data rates;
- fleet-scale collision avoidance and deorbit compliance;
- insurance capacity for large orbital compute fleets;
- ground segment, fiber backhaul, cybersecurity, and operations;
- customer workload architecture for latency-tolerant training versus latency-sensitive inference.

This is why the orbital case should be priced as a strategic option until it proves a working cost stack.

9. Where space actually wins

Space can win without beating terrestrial commodity compute.

In-orbit data processing. Earth-observation, defense, space-domain awareness, and scientific payloads benefit from processing data near the sensor. NASA's SpaceCube and HPE evidence supports this category (NASA SpaceCube; NASA Spinoff).

Defense, intelligence, and sovereignty. Certain workloads value physical separation, orbital vantage point, denial resilience, or sovereign architecture more than lowest cost per watt.

Grid-constrained premium compute. If a buyer's alternative is waiting five years for a terrestrial interconnection in a constrained market, orbital compute may price as schedule insurance. That is not a commodity-cost win; it is a scarcity product.

Space-to-space services. Lunar, cislunar, station, satellite-servicing, and autonomous spacecraft operations could value compute where terrestrial round-trip dependence is operationally costly.

Brand and sustainability optics. Space-based solar-powered compute will attract attention. But brand value is not a substitute for cost parity, thermal closure, deorbit compliance, or refresh economics.

The workload distinction is the hinge. AI training is more latency-tolerant and therefore more plausible for orbit if data ingress/egress can be managed. Inference serving terrestrial users is latency-sensitive and tightly integrated with terrestrial networks, making orbit less attractive unless the service is specialized, cached, or serving space-native users.

10. What actors should do differently

Hyperscalers should treat orbital compute as an option portfolio, not a substitute for terrestrial infrastructure. The core infrastructure program should remain staged terrestrial campuses in power-rich markets, with early transformer reservations, utility cost-sharing, BTM/bridge power where politically viable, direct-to-chip and dry/hybrid cooling, reclaimed-water strategies, and flexible-load design. Orbital pilots should focus on workloads where orbit has structural value: edge processing, defense, sovereignty, and space-native services.

Infrastructure investors should underwrite orbital compute like FOAK infrastructure, not like a data-center lease-up. The relevant questions are not only launch price and capex; they are spectrum, debris, insurance, refresh, manufacturing yield, launch cadence, and customer willingness to pay for non-cost attributes. A filing-stage constellation should not receive the same risk treatment as a permitted, utility-served, terrestrial powered shell.

Launch providers should stop selling orbital compute primarily as a launch-cost story. The investment case improves when launch providers show integrated module costs: watts per kilogram, radiator kilograms per kilowatt rejected, deployed thermal area, replacement cadence, fairing packing efficiency, launch manifest reliability, and deorbit cost.

Utilities and grid operators should assume AI load growth remains a terrestrial problem even if orbital compute succeeds in niches. Large-load tariffs, direct cost assignment, long-term

commitments, flexible-load credits, grid-enhancing technologies, and secondary-market transmission planning are the practical response. DOE's grid-enhancing technology materials and VPP work show that there is still room to expand effective grid capability (DOE GET; DOE VPP).

Regulators should avoid both extremes. Terrestrial regulators should not use data-center impacts as a reason to freeze economically valuable infrastructure; they should require transparency on water, power, generators, transmission, and ratepayer risk. Space regulators should not treat orbital compute as just another broadband constellation; compute fleets could multiply spectrum, debris, power, brightness, and launch-cadence concerns.

11. Outlook and verdict

The space-based-compute bull case is strongest when it argues that terrestrial constraints are real. It is weakest when it assumes those constraints are static and that orbital constraints are merely engineering details.

The correct comparison is not orbit versus a delayed Ashburn interconnection. It is orbit versus a global terrestrial frontier that can move to power-rich markets, build dedicated transmission, reserve transformers, use large-load tariffs, deploy BTM generation, add batteries, optimize transmission, shift workloads, improve PUE/WUE, refresh silicon, and use modular construction.

The answer to the central question—what best satisfies power, schedule, connectivity, and cost priorities—is therefore:

1. Best overall for general AI compute: terrestrial frontier data centers. They offer the best combination of power access, construction maturity, network connectivity, hardware refresh, repairability, and cost.
2. Best strategic hedge: limited orbital compute pilots. These are justified for edge processing, defense, sovereignty, space-native workloads, and learning value—not as near-term commodity replacements.
3. Not yet bankable as a commodity substitute: GW-scale orbital data centers. Starcloud and SpaceX filings are important, but filings are not deployed capacity, and announced satellite counts are not usable compute watts.

The single biggest risk to this recommendation is not that terrestrial power becomes easy. It will not. The biggest risk is that SpaceX or a similar vertically integrated actor compresses launch cadence, spacecraft manufacturing, power, thermal rejection, optical networking, and deorbit into one industrial machine—and proves a cost per usable compute watt close to terrestrial before 2040. Until that happens, orbital compute should be priced as a premium strategic product, not the next default hyperscale platform.

Source notes and open research questions

This analysis uses public agency documents, public filings, national-lab reports, reference-class project evidence, and transparent scenario modeling. Several figures are deliberately treated as contested or preliminary: the full terrestrial 1 GW capex stack, orbital mass per usable compute watt, spacecraft manufacturing cost at GW scale, Starship-style launch cost claims, and the current status of less-documented private orbital-compute actors. Where a scenario number depends on model judgment rather than a directly quoted public price, the assumption is stated in the relevant section—for example, the terrestrial IT stack is modeled as accelerators, servers, storage, networking, racks, and IT-side electrical integration, while the orbital mass stack is modeled as the full compute, power, thermal, bus, communications, shielding, redundancy, deorbit, and ground-segment system. The most important open research questions are:

- Can any orbital actor publish measured watts of usable compute per launched kilogram, including power, thermal, bus, comms, redundancy, and deorbit mass?
- Can a flight-rated accelerator refresh cycle approach terrestrial GPU refresh economics?
- What spectrum and optical-backhaul architecture supports hyperscale AI data movement without becoming the bottleneck?
- What is the demonstrated radiator mass per kilowatt rejected at relevant temperatures for AI hardware?
- What insurance market and debris-compliance regime applies to very large compute constellations?
- What customer workloads are valuable enough to pay for orbital strategic attributes even when terrestrial cost is lower?

Until those questions are answered with operating data rather than announcements, the terrestrial frontier remains the benchmark to beat.