McKinsey & Company

Technology, Media & Telecommunications Practice

The case for data centers in space

Starcloud CEO Philip Johnston on the potential role orbital data centers could play in meeting growing AI compute demand—and the technical and economic uncertainties that remain.

May 2026

As demand for AI compute rapidly accelerates, space-based data centers have the potential to move from concept to early deployment. In practical terms, this involves packaging servers and supporting systems into space-qualified modules, powered primarily by solar energy, managing tempo, and connecting back to Earth through high-bandwidth communications links.

In theory, space-based systems could offer both structural advantages, such as unconstrained energy scaling and higher solar efficiency, but also the potential for cost competitiveness with terrestrial systems. Questions remain, however, about whether space-based compute can deliver that in practice, offering predictable performance, repeatable deployment, credible reliability, and sustainable, competitive economics even after accounting for launch cadence, replacement cycles, and data-movement costs.

Philip Johnston, a McKinsey alumnus and cofounder of orbital compute infrastructure provider Starcloud, believes that space-based systems could become a meaningful part of the future compute landscape. He recently spoke to McKinsey Partner Luca Bennici about how the space-based data center technology is evolving, the challenges involved, and what needs to happen for orbital data centers to become a viable complement to terrestrial infrastructure. The interview transcript has been edited for clarity and style.

Defining data centers in space

McKinsey: Let's start with the basics. What does "data center in space" mean operationally and physically?

Philip Johnston: It's a good question because the terminology causes some confusion. Our first platform, launched in November, is a roughly one-kilowatt (kW) satellite with five embedded GPUs. It essentially had data center–grade compute in orbit—we were able to train models and run inference, including a version of Gemini, directly in space.

That said, calling it a data center is a stretch. It's more correct to say it's the first satellite carrying data center–grade GPUs. The next generation, which we will launch in about a year, is a ten-kW system that is rack-scale, with multiple advanced chips and more robust infrastructure. That's where we can plausibly argue it's part of an actual data center in space.

The real step change comes with what we call Starcloud 3, planned for 2028. That's an approximately 200-kW system, about three tons, designed to fit into Starship's deployment format. At that point, whether it's in 2028 or slips to 2029 or 2030, it's effectively a space-based data center capable of handling large-scale inference workloads.

We ultimately plan to deploy a constellation—up to 88,000 satellites—which could deliver on the order of 20 gigawatts of compute capacity, primarily for inference workflows.

What's driving demand for data centers in space

McKinsey: What's the core rationale for moving data centers into space? Cost, energy, or something else?

Philip Johnston: The biggest constraint for data centers on Earth today is energy, specifically the ability to build new energy infrastructure fast enough. In some regions, even relatively straightforward large-scale solar installations can face multiyear permitting timelines and, often, local obstacles. In space, those constraints can disappear, and we face far fewer bottlenecks to scale energy production, particularly in certain orbital configurations.

Even compared to parts of the world where energy—as well as land availability—isn't such a constraint, data centers in space hold other potential advantages: First, you don't need battery storage or other backup power for nighttime use—solar generation can be near-continuous in the right orbit. And solar panels in orbit can be far more efficient—about eight times the energy output per square meter compared to Earth.

McKinsey: What about cost considerations?

Philip Johnston: Today, space-based data centers are generally more expensive, largely due to launch costs. The long-term economics depend significantly on how those costs evolve. If launch costs decline materially, potentially to around $500 per kilogram, space-based compute becomes economically competitive with terrestrial alternatives. Below that, it becomes cheaper.

Infrastructure costs are also lower in space. We don't need batteries, backup power, chillers, cooling towers, or AC/DC conversion. That could bring infrastructure costs down to under $5 million per megawatt, versus $12 million to $15 million for terrestrial systems, at least in the US.

Who and what Starcloud is targeting

McKinsey: Who are and whom do you envision as your primary customers?

Philip Johnston: The main customers are hyperscalers and emerging AI-focused neocloud providers. What's driving demand is that it's becoming harder to find new energy sources for data centers.

We see two primary models for our services. The first one is more like a cloud provider, where we sell GPU time directly to other hyperscalers. The other is more of an energy and infrastructure play, like a colocator, where we provide the power, cooling, and connectivity, while customers choose the chip architecture and can, in turn, sell to their own customers.

McKinsey: What kinds of workloads are best suited for space-based data centers?

Philip Johnston: Almost all inference workloads are viable, as long as they don't require hyper latency, like below 50 milliseconds. Humans generally don't notice latency below approximately 200 milliseconds, so most workloads are well-suited, including chat and search queries, voice agents for customer service, back-office automation, or video generation. The only exceptions are applications that require extremely low latency—things like high-frequency trading or certain types of gaming, for instance.

In the near term, we'll mostly be providing onboard inference capabilities to other spacecraft, from the public or private sector. That's an attractive early market because it supports R&D and commands significantly higher pricing than with a terrestrial customer.

Energy and sustainability for space-based compute

McKinsey: Can you clarify how you ensure continuous power generation in space without relying on storage?

Philip Johnston: Most orbits aren't suitable because satellites spend roughly half their time behind Earth's shadow, which would require batteries, similar to the challenge of harnessing solar power on Earth.

We use a specific orbit called a dawn–dusk sun-synchronous orbit. In this approach, the satellite essentially follows the Earth's terminator line and remains in continuous sunlight. That eliminates the need for large-scale energy storage and creates a highly stable power environment, arguably more stable than on Earth, since you don't have weather, humidity, or other environmental variability.

McKinsey: How do you plan to make these data centers as or more sustainable than their terrestrial alternatives, given that customers often have very stringent sustainability requirements?

Philip Johnston: Most sustainability issues with data centers are in the form of carbon emissions. In theory, even including emissions associated with the satellite launches, space-based systems could generate significantly lower emissions than the typical gas-powered data centers. Also, when space-based infrastructure reaches the end of its useful life after about five years, it is designed to burn up in the atmosphere.

Operating and maintaining data centers in space

McKinsey: Traditional data centers rely heavily on redundancy, orchestration, and security controls. How do those translate to space?

Philip Johnston: A lot of the principles are remarkably similar to distributed satellite systems. On security, we use end-to-end encryption, including encrypted workloads for sensitive use

cases in which we don't even know what we are running on the data. On redundancy, the system is inherently distributed. If one satellite goes down, its workloads are simply routed to another. That makes the system extremely resilient.

McKinsey: What happens when something breaks? You can't exactly service a satellite easily.

Philip Johnston: We likely won't be doing any kind of robotic maintenance on the first few generations of systems, so they need to operate autonomously for their full life cycle, typically around five years. That means there must be redundancy in critical systems, overprovisioning components (like solar panels) that degrade over time, and software-based fault tolerance.

It isn't that different from modern data centers. At scale, operators often don't physically repair individual components. Instead, they route around failures and handle issues in software. We take a similar approach. If something fails and can't be recovered, it's bypassed rather than repaired.

Dealing with radiation and cooling

McKinsey: How do you deal with radiation in space, especially using commercial chips?

Philip Johnston: Radiation hardening is a critical issue. Traditional space hardware uses radiation-hardened chips, but they're extremely expensive and far less powerful—at least 100 times less powerful than state-of-the-art terrestrial options.

From the beginning, our approach has been to use state-of-the-art commercial chips and adapt them for space. That requires extensive testing.

We've exposed chips to simulated five-year radiation doses using particle accelerators, including proton and heavy ion testing. That allows us to understand failure modes and design proper shielding. We now also have real-world telemetry from orbit, which gives a unique data set on how these chips really behave in space.

McKinsey: Cooling is a major issue in terrestrial data centers. How do you manage heat in space?

Philip Johnston: Cooling in space is counterintuitive. While space is cold, there's no medium for heat transfer—no air—so you can't rely on convection or conduction. Everything has to be dissipated via radiation.

That's governed by the Stefan–Boltzmann law, which states that heat dissipation is proportional to the fourth power of absolute temperature. Small increases in radiator temperature significantly improve efficiency.

Our approach is threefold: circulate coolant through the chips, transfer heat to radiators, and emit that heat as infrared radiation.

The trade-off is surface area. For example, a system like Starcloud 3 might require solar panels covering the equivalent of several tennis courts, plus added radiator surface area to dissipate heat effectively.

The key is designing radiators that are lightweight and cost-efficient—which we've achieved at a fraction of the cost and mass of legacy systems like those on the *ISS* [*International Space Station*].

Market size and long-term vision

McKinsey: How big do you think this market could become, and what adoption curve do you expect?

Philip Johnston: If you look far enough out, the potential scale of compute demand is very large, with the upper bound essentially the total energy output of the sun.

More practically, because of energy demand, within the next decade, I expect most new compute capacity to be deployed in space. That could translate into a market of about $1 trillion per year in capital expenditure for orbital data center build-out.

The driver is simple: AI demand is growing rapidly and will eventually account for the vast majority of global compute. That trend doesn't plateau—it approaches 100 percent. If that growth continues, it could place increasing pressure on terrestrial infrastructure.

McKinsey: From a long-term investor point of view, what creates a durable competitive advantage in this area?

Philip Johnston: There are a few things. Fundamentally, you need to provide a better service—this approach should have lower energy costs and better reliability. This should also be a sticky customer base; it's a bit like with cloud service providers, where switching costs are high because the risk of losing a week or two of customer business far outweighs any potential savings from changing vendors.

There may also be scale advantages for early, fast movers, especially in areas such as system design and operational experience. Finally, IP [intellectual property] and engineering—we're already building a significant patent portfolio. For example, there are only so many ways you can build a radiator for this. We've tried all the ways that don't work, and we've patented all the ways that we believe do work.

Competition from big, established players like SpaceX is real. They have advantages in launch capabilities and execution. But we believe we will differentiate with bare metal and flexible infrastructure, allowing customers to choose their own hardware and workloads and sell them to their customers.

Even if we have a higher cost base than SpaceX, we still believe we will have a lower cost base than other hyperscalers. If they eventually conclude they need to quickly scale up space-based compute capacity, we aim to have a leading capability in orbit. That can position us as a partner to hyperscalers rather than as a direct competitor.

The Starcloud origin story

McKinsey: You had previously worked in consulting at McKinsey and didn't start as an engineer, so how did you end up focusing on data centers in space?

Philip Johnston: I've been interested in space for a long time, but the turning point was seeing how quickly launch costs were falling, especially with programs like Falcon 9 rideshare, and what that meant for the volume of satellite launches.

Three years ago, I randomly decided to go down to Texas one weekend to see SpaceX's Starship facility. The scale of what they're building amazed me; it was clear that access to space was about to change dramatically.

From there, I started talking to my old friend—and Starcloud cofounder—Ezra Feilden, an engineer who had spent a decade designing satellites, about what kind of new things such reduced launch costs could enable. Initially, we looked at all sorts of science fiction concepts we had read about as kids, most notably space-based solar power, but we quickly realized that you lose most of the energy in transmission to Earth. Moving the data centers to space is a much more efficient solution. That insight led the two of us, along with our third cofounder, Adi Oltean, to Starcloud.

Philip Johnston, a McKinsey alumnus, is CEO and cofounder of Starcloud. **Luca Bennici** is a partner in McKinsey's Dubai office.

Comments and opinions expressed by interviewees are their own and do not represent or reflect the opinions, policies, or positions of McKinsey & Company or have its endorsement.

This interview was edited by Daniel Eisenberg, an executive editor in the New York office.