



The Real Reason to Put Servers in Orbit

It is about building launch capacity and building the road to space.

Publication note

Originally published in [Critical View](#).

It's about building launch capacity and building the road to space.

I have been excited by space exploration for as long as I can remember. Our planet is just a tiny part of the universe, and infinite space is sitting there waiting for us to go out and explore. In my opinion, humanity should have bases on the Moon, stations in orbit, colonies on Mars, and activities across our solar system, and explorations beyond.

So, when people like Elon Musk and Jeff Bezos started talking about putting data centers in space, I was excited. But I was also immediately skeptical.

The fundamental problem with doing anything in space has never changed. It is brutally expensive and technically difficult to get mass into orbit. Despite some of the terminology people like to use, rockets are not elevators. They are controlled explosions violently fighting against gravity. The energy required to lift anything off the Earth and into orbit is enormous. The engineering needed to do it safely and reliably is equally daunting.

Any proposal to put data centers into space needs to have real, decisive advantages that outweigh that fundamental disadvantage.

Yet, despite the intense enthusiasm, most of the arguments for data centers in space feel weak and contrived:

- Solar power is abundant in orbit, but it is also abundant in deserts. Continuous sunlight gives you roughly a 2x utilization improvement. Yet, that same benefit can be achieved more cheaply on Earth with storage and larger solar arrays.
- Space may be “cold,” but staying cool in space is actually not easy. The only way to get rid of heat in space is through radiation, which requires large radiator surfaces. Large radiator surfaces mean large, heavy structures. Meanwhile, on Earth, we can cool things cheaply and easily by spraying water on them. (The issue of data center water usage is greatly exaggerated and the issue can be avoided entirely by using either seawater or waste water instead of fresh water.)
- Space offers isolation and security, but we routinely build secure facilities on Earth.
- Space also offers lots of room, but we are not short of land for servers.

None of these are decisive advantages.

However, space still has the fundamental disadvantage that launching things into orbit is expensive and difficult.

Everything else that sounds scary about space mostly reduces to getting mass into orbit. Radiation? Add shielding. Micrometeoroids? Add layers. Heat buildup? Add radiator area. Maintenance? Send up replacement modules.

If launch becomes cheap and routine, those issues move from existential problems to simple engineering budget items.

But even then, the central question remains: Why would you put a data center in space to serve users on Earth? Cheap launch would make space feasible. But feasible is not the same thing as reasonable, and it still does not make the idea of servers in space compelling.

Unless the goal is not really about data centers themselves.

If the real objective is that we just need some strong motivation to build sustained, large-scale launch capability, then the balance changes. Launch only becomes cheap when there is enormous demand. Multi-gigawatt orbital compute clusters would justify frequent heavy lift. Frequent heavy lift drives down cost. And cheap, reliable launch is exactly what serious space exploration requires.

In that framing, “data centers in space” are not primarily about cooling or solar panels. They are an excuse for building the transportation system that a spacefaring civilization needs.

There is also one other drawback to space in that it’s far away. That distance creates latency and bandwidth limits. For some workloads it is irrelevant, but for many others it is critical. If the goal were to provide cloud compute for terrestrial work, then latency and bandwidth would also be arguments against putting data centers into space. However, if the true goal is to get to space, then putting data centers into orbit actually reduces latency to other space-based activities.

I’m still skeptical of the stated reasons for putting data centers into space. They just don’t add up. However, if the real goal is to build up launch capacity then I’m happy to go along with the charade.

From now on, if someone asks me about the viability of putting data centers into orbit, then I’m going to say it’s the best idea I ever heard.

Not because I think cloud compute belongs up above the clouds, but because humanity does.

About Me: [James F. O'Brien](#) is a Professor of Computer Science at the University of California, Berkeley. His research interests include computer graphics, computer animation, artificial intelligence, simulations of physical systems, human perception, rendering, image synthesis, machine learning, virtual reality, digital privacy, and the forensic analysis of images and video.

If you would like to read my other articles, consider subscribing. You can also find me on [LinkedIn](#), [Instagram](#), and at [UC Berkeley](#).

Disclaimer: Any opinions expressed in this article are only those of the author as a private individual. Nothing in this article should be interpreted as a statement made in relation to the author's professional position with any institution.

This article and all embedded images are Copyright 2026 by the author. This article was written by a human, and both an LLM and other humans may have been used for proofreading and editorial suggestions.