Why we should train AI in space

Ezra Feilden PhD, Adi Oltean, Philip Johnston

Introduction

To keep pace with AI development, vast new data centers and many gigawatts of new energy projects to power them will need to be deployed around the world. At the same time, electrical utilities are being hit by a tidal wave of new demand from the electrification of industry, transport, and heating. Electricity demand may triple in the coming years as a result[1], but utilities in the Western world, hampered by planning restrictions, are not equipped for change at the required pace and scale. Without rapid adaptation, the upcoming energy crunch will hinder AI development. This issue has been flagged by multiple thought leaders in 2024:

> *"We still don't appreciate the energy needs of this technology…there's no way to get there without a breakthrough…we need fusion or we need radically cheaper solar plus storage or something" -* ***Sam Altman***

> *"We have silicon shortage today, a voltage step down transformer shortage probably in about a year, and then just electricity shortages in general in about two years" -* ***Elon Musk***

> *"We would build out bigger clusters than we currently can if we could get the energy to do it" -* ***Mark Zuckerberg***

> *"The amount of power to run compute by 2045 will be the base power of the planet right now. The drain on resources is so high, you need to put that compute in space and use the power of the sun…that's a really good use of space to help save the planet" -* ***Tom Mueller, employee #1 at SpaceX***

> *"The results of the European Commission's ASCEND study confirm that deploying data centers in space offers a more eco-friendly solution for hosting and processing data."* ***- Christophe Valorge, CTO at Thales Alenia Space.***

Aside from energy considerations, there are several other compelling reasons why Earth-based data centers do not scale well or sustainably to gigawatt (GW) sizes. For reference, large hyperscale data centers today reach 100 megawatts (MW), with some plans to approach 1 GW.[2] If the world is to continue scaling up these clusters to achieve artificial general intelligence (AGI) at the current pace, a new approach is necessary. Shifting GW scale data centers from Earth to space is a novel way to manage such a transition. While the challenges to these spacecraft will be substantial, working from first principles, Lumen Orbit has developed a

range of concept designs and has not found any insurmountable obstacles. With new, reusable, cost-effective heavy-lift launch vehicles such as Starship and New Glenn set to enter service, combined with the proliferation of in-orbit networking, the timing for this opportunity is ideal. Lumen Orbit, Inc. is at the forefront of this development as the first company to pursue orbital data centers of this scale. Our high-level vision is outlined below.

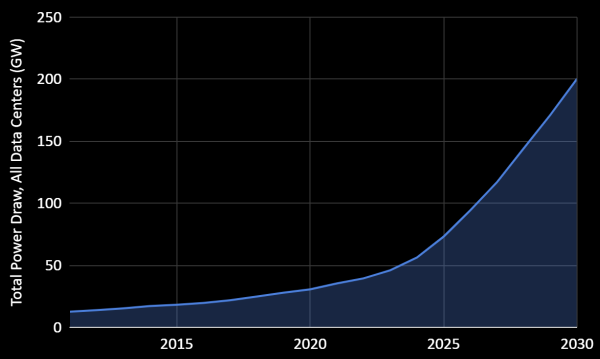

Figure 1. Global data center power demand trend and forecast. Source derived from Semianalysis. Note this is a base case, some estimates are significantly more aggressive, see Leopold Aschenbrenner.

Why Data Centers in Space?

Orbital data centers offer several fundamental benefits compared to their terrestrial counterparts, especially when scaled to GW sizes. Significant operational cost savings can be achieved by using inexpensive solar energy without the limitations of terrestrial solar farms discussed below. Orbital data centers can leverage lower cooling costs using passive radiative cooling in space to directly achieve low coolant temperatures. Perhaps most importantly, they can be scaled almost indefinitely without the physical or permitting constraints faced on Earth, using modularity to deploy them rapidly. All of this will have a net benefit on the environment - a recent study by the European Commission concluded that orbital data centers will significantly reduce greenhouse gas emissions from grid electricity and eliminate fresh water usage for cooling.[3]

Each of these benefits is detailed below and considered against the challenges and additional costs associated with deploying and operating this infrastructure in space.

Reduced Operating Expenses

Data centers in space can utilize high-intensity 24/7 solar power unhindered by day/night cycles, weather, and atmospheric losses (attenuation). This enables orders of magnitude lower marginal energy costs, resulting in drastic operating cost savings versus their terrestrial counterparts.

The performance of power plants are compared by their peak output and their “capacity factor” as follows:

$$Capacity\ factor = \frac{Power\ Output\ power\ averaged\ over\ one\ year}{Peak\ power\ output}$$

Terrestrial solar farms in the US achieve a median capacity factor of just 24%[ref], while solar projects in temperate regions such as northern Europe typically achieve capacity factors under 10%. The majority of terrestrial solar farms' generating potential is reduced by suboptimal sun position and losses due to the atmosphere and weather. A capacity factor >50% is impossible on Earth due to the day/night cycle alone. By contrast, the capacity factor of our proposed space-based solar array is greater than 95%, with no day/night cycle, optimal panel orientation perpendicular to the sun's rays, and no effects from seasons or weather. Additionally, the peak power generation will be ~40% higher than terrestrial solar farms as the atmosphere attenuates and scatters solar radiation, even on a clear day. Therefore a given solar array in space will generate over 5 times the energy as the same array on Earth. This means that it is possible to generate extremely low-cost solar energy in space. Assuming a 40 MW data center per $5m launch (see launch section below),[4] and material cost of solar cells at $0.03 per watt,[5] all amortized over 10 years, we will be able to offer an equivalent energy cost of ~$0.002/kWh. For comparison, the US, UK, and Japan typically achieve average wholesale electricity costs of $0.045/kWh[6], $0.06/kWh[7] and $0.17/kWh[8], respectively. Orbital data centers can therefore offer energy 22 times lower cost than today's energy prices. The orbital data center concept shares some similarities with space-based solar power plants but is not limited by the most challenging aspect of space-based solar - the transmission of generated power back to Earth's surface.

To achieve this performance, the effects of ionizing radiation, UV, and the thermal coefficient (reduced efficiency at high temperatures) of the solar cells need to be appropriately mitigated and balanced over the lifetime of the data center. Each of these effects can impact the output of the cells. However, with appropriate cell selection and array design, degradation rates of just 0.15% per year have been demonstrated.[9]

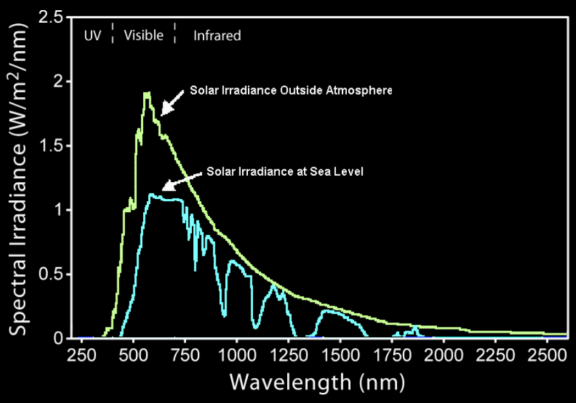

Figure 2. Solar irradiance in space vs on earth's surface, showing atmospheric attenuation of the sun's rays. Orbital data centers have access to ~40% higher solar irradiance. Source.

Additionally, deep space is cold, which is accurate in that the "effective" ambient temperature is around -270°C, corresponding to the temperature of the cosmic microwave background. An object in high orbit will naturally reach this temperature if it is shielded from the sun and Earth's reflected light. To use deep space as a heatsink to dissipate waste heat, a deployable radiator is needed. A 1m x 1m black plate kept at 20°C can radiate about 850 watts to deep space, which is roughly three times the electricity generated per square meter by solar panels. As a result, these radiators need to be about one-third the size of the solar arrays, depending on the radiator configuration. In space, we can use simpler and more efficient cooling architectures than energy-intensive chillers, which are used to achieve low temperatures in terrestrial data centers. We estimate we can achieve comparable PUE to state-of-the-art hyperscale terrestrial data centers. Additionally, orbital data centers in certain orbits experience virtually no fluctuation in "ambient temperature" (beyond ~0.2% variation in solar irradiance) and exist in a highly stable thermal and mechanical environment, aiding thermal control and stability. On Earth, cooling systems must be designed for the hottest days, sometimes exceeding

45°C, leading to significant overprovisioning for average conditions. In space, there is no need for such overprovisioning.

Table 1. Cost comparison of a single 40 MW cluster operated for 10 years in space vs on land.

Cost Item	Terrestrial	Space
Energy (10 years)	$140m @ $0.04 per kWh	$2m amortized cost of solar array
Launch	None	$5m (single launch of compute module, solar & radiators)
Cooling (chiller energy cost)	$7m @ 5% of overall power usage	More efficient cooling architecture taking advantage of higher ΔT in space
Cooling (water usage)	1.7m tons @ 0.5L/kWh[10]	Not required
Enclosure (Satellite Bus/Building)	Approximately equivalent cost	
Backup power supply	$20m (commercial equipment pricing)	Not required
All other data center hardware	Approximately equivalent cost	
Radiation shielding	Not required	$1.2m @ 1kg of shielding per kW of compute and $30/kg launch cost
Cost Balance	**$167m**	**$8.2m**

Scalability

Orbital data centers unlock next-generation clusters of a scale not seen yet on Earth, with power generation well into the GW range. They can be linearly scaled nearly indefinitely without the physical and planning constraints that plague terrestrial projects of this size.[11] If current trends continue, multi-GW clusters will be required from 2027 to train the largest LLMs.[12] Consider a 5 GW cluster, which will be needed to train models like Llama 5 or GPT-6. This would exceed the capacity of the largest power plant in the US and some of the largest operational power plants in the world. These clusters are, therefore, simply not possible with today's energy infrastructure. At the same time these clusters will be essential to train next-generation AI models of the future.

To scale to gigawatts in orbit, compute modules, power, cooling and networking can be assembled together in a modular fashion. Compute modules can also be assembled with architectures that scale in 3D rather than 2D as on Earth, ensuring the cluster is as tightly coupled with as low latency within the cluster as possible (a critical property of AI training clusters). There are also potential opportunities to leverage the fact that the speed of light in a vacuum is 35% faster than in a typical glass fiber.

Speed of Deployment

In Western countries, new large-scale energy and infrastructure projects often take a decade or more to complete due to myriad permitting requirements, rights of way and utility/transmission line restrictions, and environmental reviews. These bottlenecks endanger the timeline of very large data centers and are already being felt. For example, xAI recently resorted to temporarily using MW-scale natural gas generators for their Memphis cluster as the grid was not ready to supply sufficient power.[13]

Orbital data centers avoid almost all of these roadblocks, but one key hurdle is on-orbit safety and orbital debris mitigation concerns, including decommissioning. In the US commercial spacecraft must submit an Orbital Debris Assessment Report to the relevant national regulator to demonstrate the probability of collisions with other objects is sufficiently low. Given their larger physical size, orbital data centers must be especially responsible users of low Earth orbit by ensuring highly responsive spacecraft maneuverability for collision avoidance, use of state-of-the-art space-object tracking systems, registering spacecraft ephemeris with responsible tracking databases, and coordinating with all relevant bodies.

It should be noted that the large majority of the surface area of orbital data centers will be solar arrays. Results from the International Space Station have shown that small debris collisions with solar arrays are generally passive over time.[14] Placing orbital data centers in underutilized orbits is also an effective strategy to mitigate orbital debris, which has partially driven our orbit choice detailed below. Other traditional spacecraft regulatory hurdles, such as radio spectrum availability, can be offset by using optical (laser) communications that are not currently subject to regulation and are much better suited for this type of high-data transfer application than radio frequency options in terms of throughput and security. Lastly, if best practices are followed[15] then the terrestrial astronomy community is unlikely to be impacted by orbital data centers in our selected orbit since they will only be visible at dawn/dusk where there is too much ambient light for most astronomical purposes.

The reduction in permitting constraints by moving data centers to space will save significant costs, but most importantly, the speed of deployment could be substantially faster than deployment of comparable terrestrial data centers. This also means that orbital data centers can be deployed in a more agile manner - scaling up faster if needed, with the freedom to change plans if commercial requirements change.

Design principles for orbital data centers

The basic design principles below were adhered to when creating the concept design for GW scale orbital data centers. These are all in service of creating a low-cost, high-value, future-proofed data center.

1. **Modularity**: Multiple modules should be able to be docked/undocked independently. The requirements for each design element may evolve independently as needed. Containers may have different compute abilities over time.
2. **Maintainability**: Old parts and containers should be easy to replace without impacting large parts of the data center. The data center should not need retiring for at least 10 years.
3. **Minimize moving parts and critical failure points**: Reducing as much as reasonably possible connectors, mechanical actuators, latches, and other moving parts. Ideally each container should have one single universal port combining power/network/cooling.
4. **Design resiliency**: Single points of failure should be minimized, and any failures should result in graceful degradation of performance.
5. **Incremental scalability:** Able to scale the number of containers from one to N, maintaining profitability from the very first container and not requiring large CapEx jumps at any one point.

Network architecture

When designing orbital data centers we can be guided by existing practices in designing container-based terrestrial data centers. Each container has sets of racks containing the compute and storage units, built-in networking, power and cooling infrastructure. The container is designed to dock with the main structure using a single mechanical port, allowing network, power, and cooling connectivity with the rest of the data center.[16] This port will contain the necessary operational support for reliably connecting potentially thousands of fiber pairs, high-power input voltage connectors and high-volume cooling.

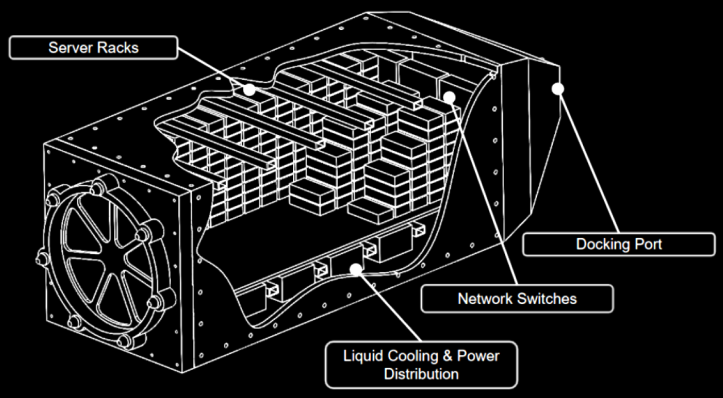

Figure 3. Compute container schematic.

Since we are focusing on AI training workloads for a space-based data center, there are a few particularities that drive the design. Space-grade data centers will be attractive for organizations that train very large AI models that could vastly exceed in size AI models trained on terrestrial data centers (limited in the amount of drawn power due to local energy supply constraints). The remaining GPU capacity can be used for other workloads, such as inference or other forms of general-purpose computing.

From the perspective of networking, training large AI models requires very low latency between all computing nodes within the data center. This implies that the containers need to be deployed physically close to each other within the data center in a tightly-connected daisy-chain-style network. We assume that all containers in a given data center should be within a few hundred meters of each other. Second, the network between all containers needs to support sufficient bisection bandwidth to efficiently train the largest AI models, which would likely consume a large fraction, if not all the data center during training. We assume that the necessary spine infrastructure, which mechanically supports all the containers, will also contain multiple layers of directories and switches appropriately sized for the workloads that the data center is designed to execute.

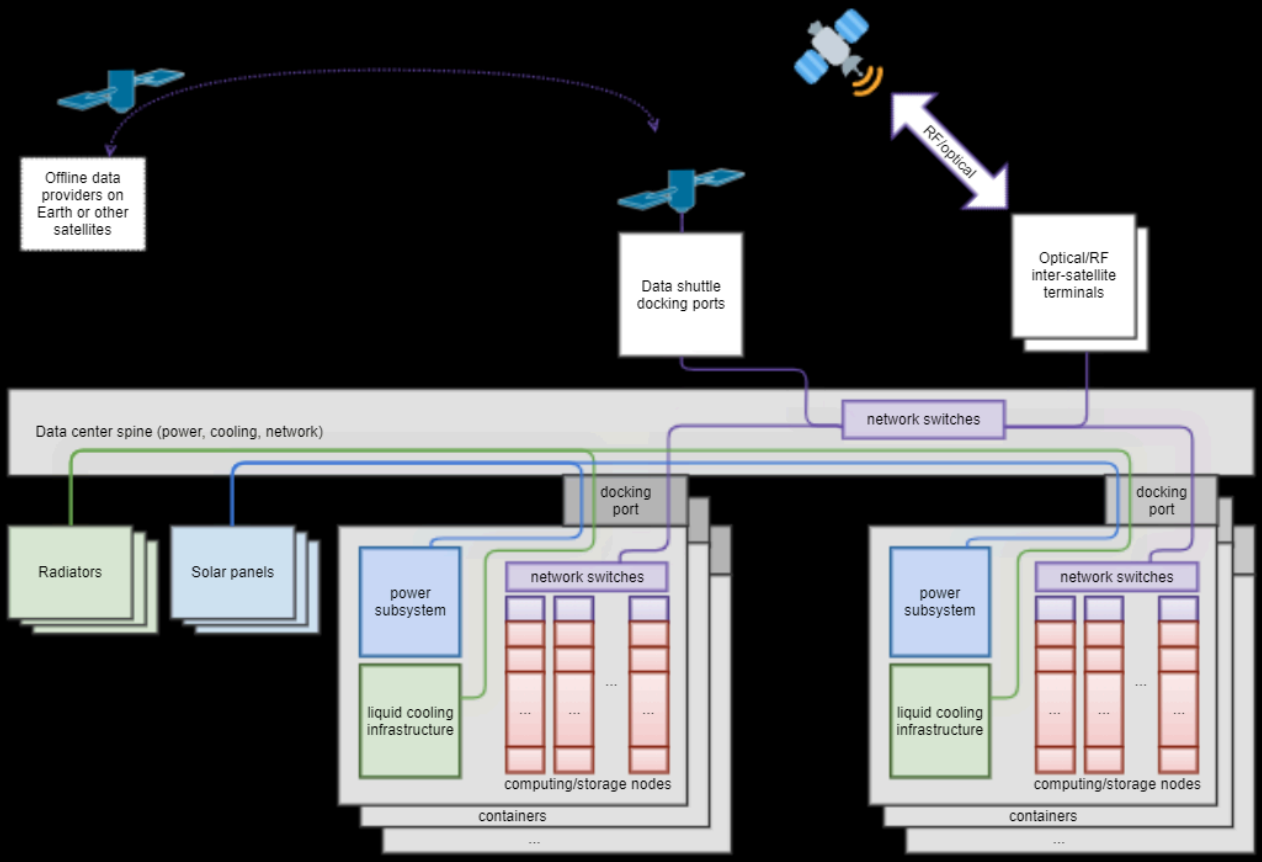

Figure 4. Orbital data center network architecture.

For connectivity, we envision using laser-based connectivity with other constellations such as Starlink, Kuiper, or Kepler. To complement optical or RF links, another method of transporting large volumes of training data would be to use data shuttles, consisting of small docking modules launched from the ground that can be used to easily transport petabytes or even exabytes of data in a single trip. This approach, popularized by Jim Gray in the '90s, was first proven to work in space by shipping 7 GB of data using an Amazon Snowcone to ISS.[17]

Physical Architecture

Power

A 5 GW data center would require a solar array with dimensions of approximately 4 km by 4 km, assuming a cell fill factor of 90% and beginning-of-life efficiency of 22% using silicon solar cells. These cells are manufactured at huge scale today (>300 GW[18] deployed in 2023, the vast majority silicon) and can cost as little as $0.03 per watt.[19] The array will be substantially smaller and lower cost than an equivalent capacity terrestrial solar farm due to the higher capacity factor and peak generation in space compared to on Earth.

To maximize the array size deployed per launch, thin film cells should be used. These cells use silicon wafers <25 μm thickness and achieve power densities >1000 W/kg,[20] allowing for highly mass and volume efficient arrays. Furthermore, these cells maintain high efficiency over their lifetime without the need for cover-glass, as

they anneal (heal) radiation damage at moderate temperatures. At these thicknesses, the cells can be folded and rolled into a very compact configuration during launch and deployed to a very large area in space.

Deployment of these arrays can feasibly be accomplished using design concepts that have been demonstrated in orbit, such as Z-fold, roll-out, or picture-frame designs. This is a core area of development within Lumen Orbit. Novel solutions are also required to ensure these structures remain controllable by the attitude determination & control system (ADCS). These systems are also in development at Lumen Orbit.

Power transfer from the arrays to the compute modules is best facilitated with high-voltage DC (HVDC) lines with appropriate electrical insulation. As on earth, the selected voltage must balance the complexity and efficiency of DC-DC converters with the gauge of copper wiring needed.

Figure 3. A data center in Sun Synchronous Orbit, showing a 4km x 4km deployed solar array and radiators.

Thermal Management

Essentially all of the power generated by the solar arrays and the remaining solar energy absorbed by exposed surfaces will need to be dissipated as waste heat. As conduction and convection to the environment are not available in space, this means the data center will require radiators capable of radiatively dissipating gigawatts of thermal load. To achieve this, Lumen Orbit is developing a lightweight deployable radiator design with a very large area - by far the largest radiators deployed in space - radiating primarily towards deep space, which has an average temperature of about 2.7 Kelvin or -270°C[21]. The thermal management system also includes mechanisms to transport the thermal load from the compute modules to the radiating surfaces. These radiators may be mechanically coupled with the deployable solar arrays. This component represents the most significant technical challenge required to realize hyperscale space data centers.

Within the compute modules, either direct-to-chip liquid cooling or potentially two-phase immersion cooling is required to achieve high power densities and a space-efficient rack setup. This is necessary to maximize the compute per launch. The principles of this cooling subsystem can mimic those of terrestrial data centers, albeit for some hardware changes, such as reducing the mass of the coolant and cold blocks. The compute modules may be either pressurized with an inert atmosphere to provide forced-convective cooling to any components that are not directly liquid-cooled, or submerged into coolant, which can also provide additional radiation shielding.

The solar arrays themselves may be passively thermally managed without the need to pump coolant to them by application of emissivity-controlling coatings to their rear surface.

Launch

The world is on the verge of a step change in launch costs, thanks to the development of several partially or fully reusable heavy-lift launchers, such as Starship, New Glenn, and Long March 9. For instance, a launch vehicle like Starship is expected to offer a launch price of around $5 million per launch long term. With a payload capacity of 100 tons to Low Earth Orbit (LEO) Sun-Synchronous Orbit (SSO), this translates to approximately $30 per kilogram. It has been suggested that costs could drop to as low as $10 per kilogram.[22] At these price points, launch costs are no longer a primary cost driver for orbital data centers.

From the perspective of networking architecture and radiation shielding, it is desirable to maximize the size of each compute container to the extent that a single container could fully occupy the launch vehicle payload bay and mass capability. This size of each container is limited only by ground test facilities and the payload capabilities of the next generation of heavy-lift launch vehicles, effectively capping each container at ~100 tons. The volume of the payload bay of these vehicles can accommodate ~300 racks at 50% capacity, with the remaining volume housing supporting systems. Assuming a power density of 120 kW per rack, equivalent to the Nvidia GB200 NVL72,[23] one launch can deploy ~40 MW of compute with rack-level mass savings. Power densities are projected to rise dramatically in the coming years, so this estimate is conservative. It is, therefore, conceivable that 5 GW of compute could be deployed with fewer than 100 launches, with a similar number of launches required for the combined solar/radiator modules of Lumen Orbit's design. These vehicles are being designed to launch up to three times per day. Therefore one launcher could conceivably launch the entire 5 GW data center in 2-3 months. As such, launch cadence will not be a bottleneck long term.

Despite this capability, a more likely scenario is a gradual buildout of the data center, using a modular design of containerised computing modules gradually assembled around a central hub with solar/radiator modules being incrementally added, radiating outwards to form a plane. This design uses just two primary structure types, allowing for economies of scale during manufacturing and reduced engineering effort.

Orbit

The choice of orbit must balance factors, including radiation, aerodynamic drag, network availability and latency, space debris, and launch accessibility. However the most important factor is continuous solar power generation, and thus a low-Earth, dawn-dusk sun-synchronous orbit (SSO) has been selected. In this orbit, the spacecraft orbits above the day/night line on Earth (known as the "terminator"). The plane of the orbit precesses around the Earth at a rate of one rotation per year. Thus the plane of the orbit remains approximately perpendicular to the direction of the sun year-round, with the spacecraft in near-continuous solar illumination. This is the only low-Earth orbit with this property. Continuous illumination is crucial as it nearly doubles the average power generation compared to orbits that see a day/night cycle, reduces fatigue from thermal cycles of the panels if the orbit is partially obscured in the shadow of Earth, and allows the data center to operate continuously without significant battery storage.

While radiation levels are low compared to many other orbits, shielding is required to reduce radiation-induced effects, including latch-up, transients, and total ionizing dose (TID) effects in sensitive components such as storage and power delivery. Note that logic devices have been shown to be resilient to radiation,[24] especially when used in AI training applications. The mass of radiation shielding scales linearly with the container surface area, whereas the compute per container scales with the volume. Therefore the mass of shielding needed per compute unit decreases linearly with container size. This effect, combined with the shielding afforded by the cooling blocks, means that radiation shielding is proportionally a much smaller concern compared to electronics on typical satellites today.

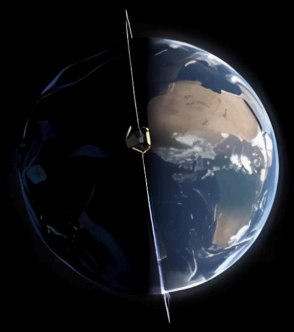

Figure 4. Satellite in a permanently illuminated dawn-dusk sun-synchronous orbit. This orbit will follow the dawn-dusk line, known as the terminator, year-round. Credit DALL-E.

Maintenance

Despite advanced shielding designs, ionizing radiation, thermal stress, and other aging factors are likely to shorten the lifespan of certain electronic devices. However, cooler operating temperatures, mechanical and thermal stability, and the absence of a corrosive atmosphere (except for atomic oxygen, which can be readily mitigated with shielding and coatings) may prolong the lifespan of other devices. These positive effects were observed during Microsoft's Project Natick, which operated sealed data center containers under the sea for years.[25] Before scaling up, the balance between these opposing effects must be thoroughly evaluated through multiple in-orbit demonstrations.

The data center architecture has been designed such that compute containers and other modules can be swapped out in a modular fashion. This allows for the replacement of old or faulty equipment, keeping the data center hardware current and fresh. The old containers may be re-entered in the payload bay of the launcher or are designed to be fully demisable (completely burn up) upon re-entry. As with modern hyperscale data centers, redundancy will be designed-in at a system level, such that the overall system performance degrades gracefully as components fail. This ensures the data center will continue to operate even while waiting for some containers to be replaced.

The true end-of-life of the data center is likely to be driven by the underlying cooling infrastructure and the power delivery subsystems. These systems on the International Space Station have a design lifetime of 15 years[26], and we expect a similar lifetime for orbital data centers. At end of life, the orbital data center may be salvaged[27] to recover significant value of the hardware and raw materials, or all of the modules undocked and demised in the upper atmosphere by design.

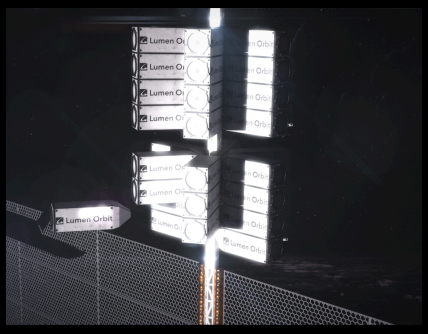

Figure 5. Using a stem and leaf design containers may be readily swapped out. Full video here.

Conclusion

Gigawatt-scale orbital data centers are among the most ambitious space projects of all time, sitting at the intersection of four trends: the drastic fall in launch costs, the upcoming electricity demand crunch, the growth in demand for large, energy-intensive GPU clusters, and the proliferation of low-cost connectivity from mega-constellations. We are convinced that orbital data centers are feasible, economically viable, and necessary to realize the potential of AI, the most important technology of the 21st century, in a rapid and sustainable manner.

References

1. https://energycentral.com/news/musk-says-triple-electricity-needed-sustainable-energy-future
2. https://www.datacenterfrontier.com/hyperscale/article/55021675/the-gigawatt-data-center-campus-is-coming
3. https://www.thalesaleniaspace.com/en/press-releases/thales-alenia-space-reveals-results-ascend-feasibility-study-space-data-centers-0
4. https://nautil.us/the-profound-potential-of-elon-musks-new-rocket-238201/
5. https://www.pv-tech.org/industry-updates/white-paper-on-182mm-wafer-based-module%EF%BC%9A-optimal-module-solution-for-achieving-lower-lcoe-at-utility-scale-photovoltaic-power-stations/
6. https://www.eia.gov/electricity/wholesale/
7. https://grid.iamkate.com/
8. https://www.renewable-ei.org/en/statistics/electricitymarket/
9. https://ntrs.nasa.gov/api/citations/20030068268/downloads/20030068268.pdf
10. https://www.ramboll.com/en-us/insights/decarbonise-for-net-zero/how-should-data-centers-be-designed-to-be-future-proof
11. https://www.networkworld.com/article/3497123/google-ireland-bid-to-build-new-data-center-rejected.html
12. https://situational-awareness.ai/racing-to-the-trillion-dollar-cluster/
13. https://www.tomshardware.com/tech-industry/artificial-intelligence/elon-musks-new-worlds-fastest-ai-data-center-is-powered-by-massive-portable-power-generators-to-sidestep-electricity-supply-constraints
14. https://www.sciencedirect.com/science/article/abs/pii/S0273117797004134
15. https://cps.iau.org/
16. https://web.eecs.umich.edu/~mosharaf/Readings/DC-Computer.pdf
17. https://aws.amazon.com/blogs/aws/how-we-sent-an-aws-snowcone-into-orbit/
18. https://ember-climate.org/insights/in-brief/2023s-record-solar-surge-explained-in-six-charts/
19. https://www.pv-tech.org/industry-updates/white-paper-on-182mm-wafer-based-module%EF%BC%9A-optimal-module-solution-for-achieving-lower-lcoe-at-utility-scale-photovoltaic-power-stations/
20. https://solestial.com/
21. https://en.wikipedia.org/wiki/Cosmic_microwave_background
22. https://twitter.com/elonmusk/status/1328770804222468097
23. https://www.semianalysis.com/p/gb200-hardware-architecture-and-component
24. https://ieeexplore.ieee.org/document/9286222
25. https://news.microsoft.com/source/features/sustainability/project-natick-underwater-datacenter/
26. https://spacese.spacegrant.org/uploads/images/ISS/ISS%20SE%20Case%20Study.pdf
27. https://virtussolis.space/blog/end-of-life-and-salvage-for-a-solar-power-satellite