To Boldly Go: The Case for Space Datacenters

Space DC Total Cost of Ownership Explained. Unpacking constraints from Terrestrial DCs and Chip Production. Space-Earth Parity in the late 2030s, Space DCs could start to be viable even sooner.

DANIEL NISHBALL, PRANAV MYANA, ELLIE HOLBROOK, AND 7 OTHERS

JUN 03, 2026 · PAID

Everyone has been talking about datacenters in space. Interviews given by Elon Musk in the past few months have spent lots of time on orbital compute:

"Five years from now, my prediction is we will launch and be operating every year more AI in space than the cumulative total on Earth... I would expect to be at least, sort of five years from now, a few hundred gigawatts per year of AI in space and rising."
- Elon Musk on Dwarkesh Podcast, February 2026

Furthering space-based compute was also one of the stated motivations behind the merger of xAI into SpaceX (as a 'reorganization of entities under common control'), and is a key part of SpaceX's plans to go public, as stated in their S-1 filing on 20 May 2026.

"Our goal over time is to launch 100 gigawatts of compute to space each year. If operated continuously, the generation resources used to support 100 gigawatts of compute could generate approximately one-fifth of the annual power production in the United States, which was 4.4 thousand terawatt hours in 2025... We expect space-based compute to massively increase AI compute scale, while also improving token economics."
- SpaceX, S-1 Filing, May 2026

As expected, many part-time prognosticators in the Substack-verse have emerged from the woodwork to weigh in on the concept. Some articles bring up insightful points, but there are more than a few that are built upon ideas that fly in the face of science.

A few casual arguments made in favor of space datacenters include the following:

1. Space can provide free solar energy 24 hours a day
2. Cooling is "free". Some erroneously point to space being cold as a key positive
3. Communications latency in space is low as you're just sending light through a vacuum
4. There is no need for permitting in space... so far...

Many of these points sound like they hold merit on the surface, but a deeper analysis of each apparent advantage reveals a far more complex story.

While we think that it is possible that space datacenters could scale one day, deploying orbital compute using today's technology currently costs several times more than deploying terrestrial compute. Achieving Space-Earth cost parity will require significant engineering work, material science breakthroughs and cost scaling progresses and will still take years to achieve. There are also important reliability and servicing obstacles to overcome - for instance - how GPU servers will recover from faults that require human intervention, effectively shielding accelerators from radiation, among many others.

When we deploy compute in space, it won't be because of the four superficial reasons we have cherry-picked above. Rather, **Space-based datacenters make sense in the world where AI demand well exceeds all of the four layers of terrestrial datacenter supply that we will introduce below**. For Space datacenters to step up to this call - it is a necessary condition that major space datacenter cost items like radiators, solar arrays and launch costs decline considerably, and that a number of key operational obstacles are overcome.

The four layers of incremental power supply for terrestrial datacenters include:

1. Grid-connected supply,
2. Converted bitcoin miners and powered land,
3. Behind the meter generation, and finally,
4. Industrial capacity and manpower to build further power infrastructure.

A necessary condition for AI related IT equipment demand to reach levels exceeding terrestrial datacenter supply is for there to be enough chip fabrication capacity to fulfill this demand in the first place, before we even discuss datacenters! We wrote about this in great detail in our recent article on the Great AI Silicon Shortage, where we concluded that the industry has moved from a power-constrained to an accelerator-constrained regime. Available datacenter capacity and power now exceed AI compute demand, but TSMC's N3 wafer capacity and HBM supply cannot keep pace with the pace of accelerator deployments. This means that today, and for the next few years, chip manufacturing will be the global constraint before we even worry about supply for these four layers.

The chip constraint forms a **separate fifth layer of supply - Semiconductor Production, and it is a "universal" constraint on all chip deployment, whether deployed on Earth or in Space.** Users of our AI Space Datacenter TCO Model can see how this constraint applies well into the future, and under what scenarios regarding chip manufacturing capacity addition that Semiconductor Production may not be the constraint.

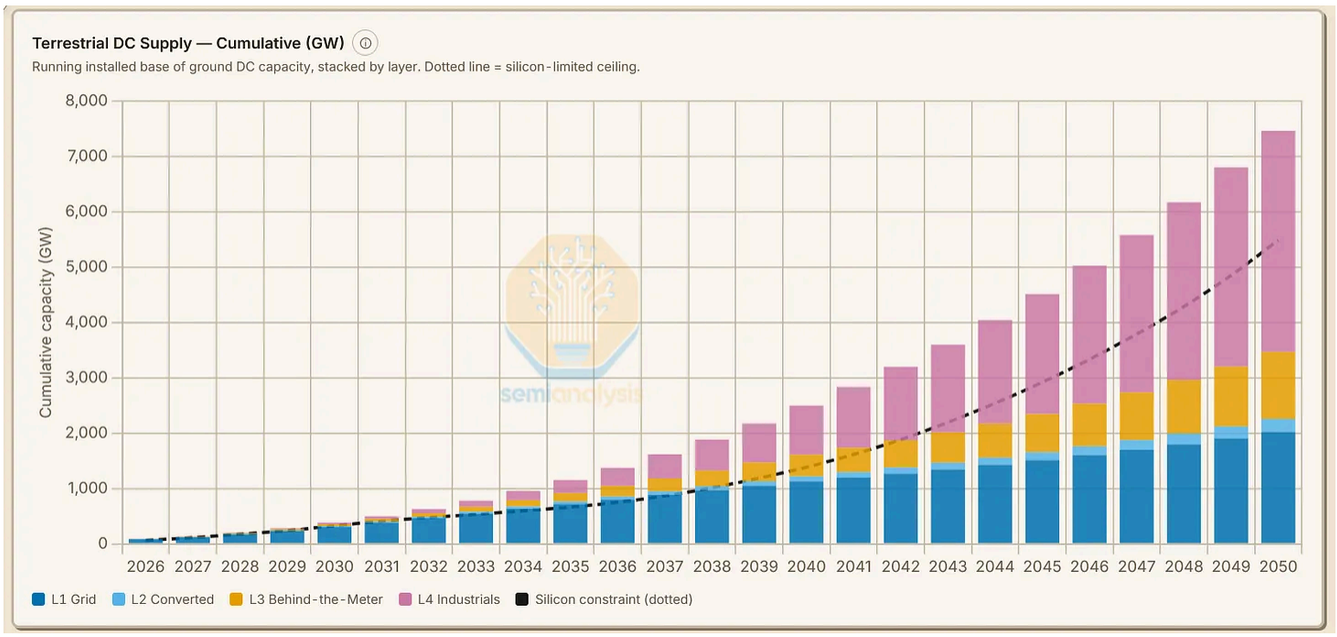
Elon Musk is clearly well aware of this constraint, and it is the impetus behind his Terafab Initiative. The AI Space Datacenter TCO Model also includes knobs and sliders for users to tune to test out various Terafab scenarios.

Framing the Space Datacenter Debate

Our various industry models such as the Accelerator Model, the Foundry Industry Model and WFE Models illustrate the aforementioned chip tightness. Meanwhile our AI Datacenter Model forecasts accelerating incremental datacenter additions in 2027 and 2028. Thus, datacenter capacity addition will run ahead of chip constraints in the next few years until fab capacity additions accelerate to catch up. Our suite of industry models will only forecast such wafer fab and datacenter capacity additions once such plans are confirmed.

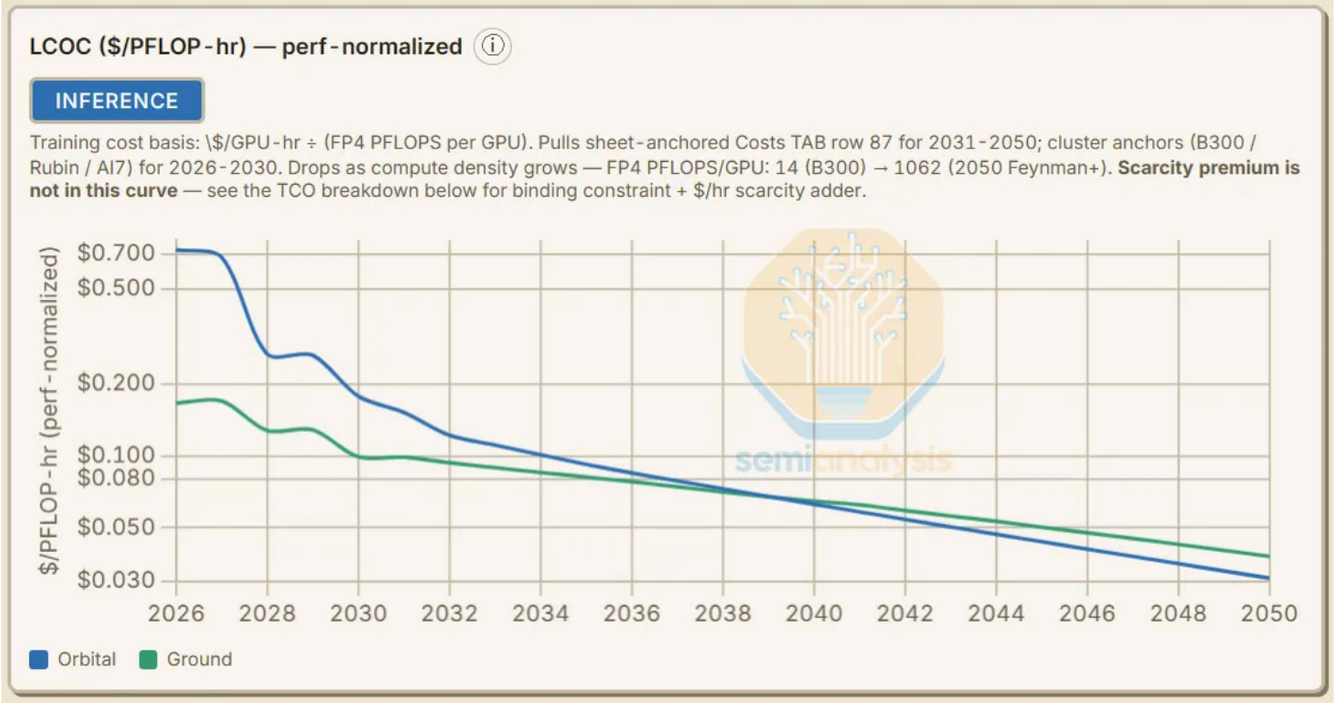
However, the world in which AI demand is so overwhelming as to exceed the already formidable datacenter capacity additions is a world with no time for half measures. As such, our AI Space Datacenter TCO Model base case departs from our industry models to reflect this world, assuming accelerating incremental datacenter capacity additions and a meaningful step up in the pace of chip fab capacity addition. It is a world where all the stops are pulled out and many obstacles from gas turbine availability to EUV tool production constraints are overcome because clear long-term AI end use ROI justifies enough capital investment to overcome them.

The below chart illustrates what this world could look like - with incremental datacenter capacity additions eventually in the hundreds of GW annually, though adding chip capacity will still be more difficult than adding datacenter capacity:



Source: SemiAnalysis AI Space Datacenter TCO Model

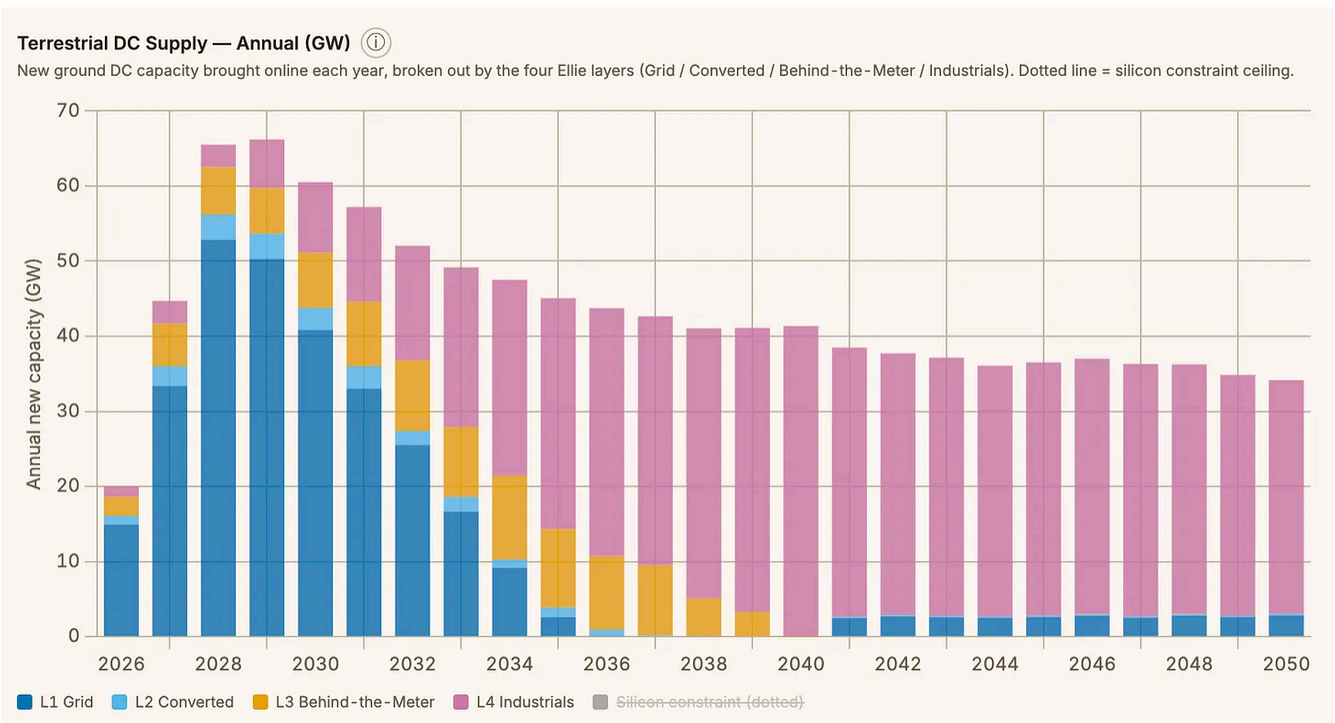
In such a base case, our model shows that the cost difference between space and terrestrial datacenters starts at more than 4x in 2026 before narrowing to parity in ~2040, with levelized costs of compute in space declining below terrestrial thereafter. **By the early 2030s, space datacenters could be only ~30% more expensive than Earth-based datacenters, opening the door to the first scaled space datacenters as soon as the turn of the decade!**



Source: SemiAnalysis AI Space Datacenter TCO Model

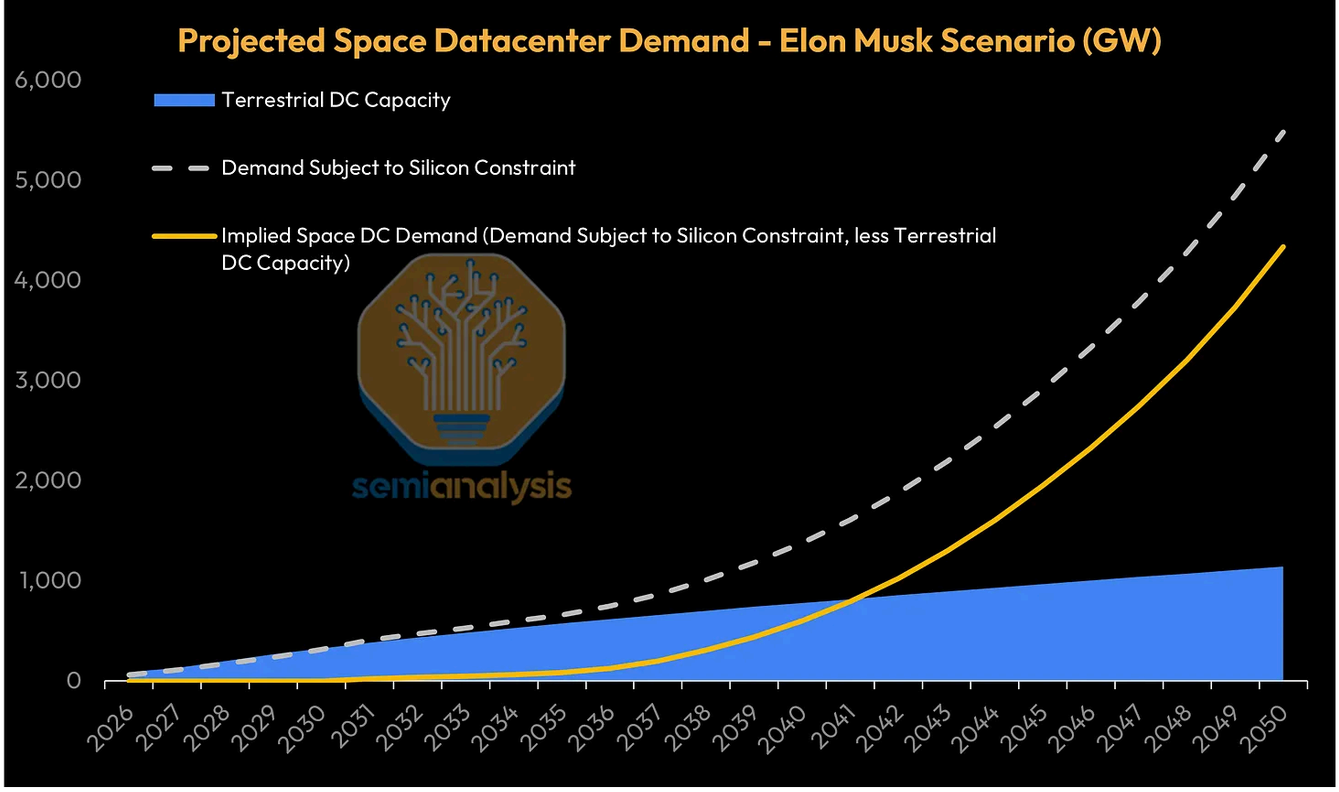
Space-Earth datacenter cost parity opens the door, but in our base case, there is still ample terrestrial capacity - so going into space is a matter of preference and optimization rather than necessity. But if regulatory and capacity bottlenecks starve terrestrial datacenter capacity - space becomes a necessity. This scenario is the basis for our model's Elon Musk Scenario.

In the Elon Musk scenario, terrestrial datacenter incremental capacity addition peaks out in 2028 and remains low for decades while chip production expansion marches ahead. In this scenario, Space becomes the only alternative for scaled AI datacenter deployments.



Source: SemiAnalysis AI Space Datacenter TCO Model

In the Elon Musk scenario, the space datacenter TAM could easily reach high hundreds of GW of incremental capacity per year.



Source: SemiAnalysis AI Space Datacenter TCO Model

Our Total Cost of Ownership Framework is the foundation that allows calculation of levelized cost of compute and it powers the AI Space Datacenter TCO Model. In the rest of this introduction, we will briefly explain the TCO framework and key conclusions from our AI Space Datacenter analysis.

Analyzing Space vs Earth Datacenters using our Total Cost of Ownership (TCO) Framework

In a manner similar to our standard AI Cloud TCO model, we split out various cost categories for space and terrestrial deployments, namely IT cluster capital cost, datacenter capital cost (including launch costs), and operating cost.

Headline TCO Findings using Present Costs and Technology

A 30.5kW B300 cluster (16 GPUs across two servers) deployed in 2026 has a total program capital cost (IT cluster capital cost plus datacenter capital cost) of $4.1M for a space deployment and $1.4M for terrestrial deployments. We use B300s and contemporary mainstream GPU as references for TCO analysis for the next few years, **but it is much more likely that smaller, efficient and specialized chips akin to Tesla's FSD chips will actually be deployed**.

Transformed into a levelized monthly ownership cost incorporating respective WACCs and useful lives and adding monthly operating costs, we see a total monthly cost of ownership of $100,925/month for space deployments vs $27,724/month for terrestrial deployments.

Space datacenters are more costly because of a larger upfront capital cost of deployment, with the largest driver being launch costs at $1.6M out of the total $3.1M datacenter capital cost. The cost difference is even starker when considering monthly levelized datacenter costs - because space datacenters are expected to have a useful life of only 5 years vs the 15 years for Earth-based datacenters, monthly levelized datacenter capital costs are a whopping 18x higher than for terrestrial datacenters!

AI Cloud Total Cost of Ownership Summary: Space vs Terrestrial (Extended Prices)

	Unit	B300 1200W (Orbital)	B300 1200W
Site		Space	Terrestrial
Customer Profile		Hyperscaler	Hyperscaler
Year		2026	2026
Program Power	W	30,565	30,565
Power per GPU	W	1,910	1,910
GPUs	GPUs	16	16
GPU per Server	GPUs	8	8
IT Cluster Capital Cost	USD	$980,882	$986,158
Datacenter Capital Cost[1]	USD	$3,086,332	$382,061
Total Program Capital Cost	**USD**	**$4,067,215**	**$1,368,219**
Weighted Average Cost of Capital	%	15.0%	10.3%
Datacenter Useful Life in Years	Years	5	15
Monthly IT Cluster Capital Cost of Ownership	USD	$23,335	$21,099
Monthly Datacenter Capital Cost of Ownership[1]	USD	$73,424	$4,176
Monthly Operating Cost of Ownership	USD	$4,167	$2,449
Total Monthly Cost of Ownership	**USD**	**$100,925**	**$27,724**

1. Includes launch cost.

Source: SemiAnalysis AI Space Datacenter TCO Model

GPU rental pricing is typically quoted on a $/hr/GPU basis - and so we also present the above costs on a per hour basis. Using today's technology, total cost of ownership on a GPU-hr basis for a B300 GPU is estimated at $8.64/hr/GPU for a space-based deployment, compared to $2.37/hr/GPU for a terrestrial deployment.

In addition to total cost of ownership (TCO) in $/hr/GPU, we also look at Levelized Cost of Compute (LCOC). The difference between TCO and LCOC is that LCOC reflects the net cost of compute needed to meet a certain SLA, given expected cluster reliability. LCOC will be higher than TCO because operators will need to account for radiation availability (i.e. compute availability that is temporarily affected by solar radiation), as well as additional GPUs required (i.e. provision for redundancy given GPU failures that cannot be replaced or repaired).

For terrestrial datacenters, the radiation availability and SLA requirement result in a gross up of about 5% on top of TCO of $2.37/hr/GPU to an LCOC of $2.49/hr/GPU. For space, this means a much larger 26% gross up on top of the $8.64/hr/GPU TCO to reach an LCOC of $10.91/hr/GPU.

AI Cloud Total Cost of Ownership Summary: Space vs Terrestrial (per GPU-hr)			
	Unit	B300 1200W (Orbital)	B300 1200W
Site		Space	Terrestrial
Customer Profile		Hyperscaler	Hyperscaler
Year		2026	2026
IT Capital Cost of Ownership	USD/hr/GPU	$2.00	$1.81
Datacenter Capital Cost of Ownership	USD/hr/GPU	$6.29	$0.36
Operating Cost of Ownership	USD/hr/GPU	$0.36	$0.21
Total Cost of Ownership, Pre-SLA	**USD/hr/GPU**	**$8.64**	**$2.37**
Datacenter Capital Cost as % of TCO	*%*	*72.8%*	*15.1%*
Radiation Availability	%	95%	100%
Additional % GPUs needed for 99% SLA	%	20%	5%
IT Capital Cost of Ownership, Post-SLA	USD/hr/GPU	$2.52	$1.90
Datacenter Capital Cost of Ownership, Post-SLA	USD/hr/GPU	$7.94	$0.38
Operating Cost of Ownership, Post-SLA	USD/hr/GPU	$0.45	$0.22
Levelized Cost of Compute (LCOC)[1]	**USD/hr/GPU**	**$10.91**	**$2.49**
Marketed PFLOPS (FP4)	PFLOPS/GPU	15.0	15.0
Inference Throughput[2]	Tok/s/GPU	5,133.0	5,133.0
LCOC per PFLOP-hour	$/PFLOP-hr	$0.73	$0.17
LCOC per B Tokens	$/B tokens	$590.67	$134.87

1. Levelized Cost of Compute; refers to total cost of ownership after adjusting for radiation availability and additional GPUs needed.

2. DeepSeek R1 FP4. Uses 8k input, 1k output and 1k input, 1k output tokens 50/50 mix, 100 interactivity.

Source: SemiAnalysis AI Space Datacenter TCO Model

All of the above TCO calculations including a further more detailed space cost breakdown are available for all years from 2026 to 2050 in our SemiAnalysis AI Space Datacenter TCO Model.

When Will Space and Terrestrial Costs Reach Parity?

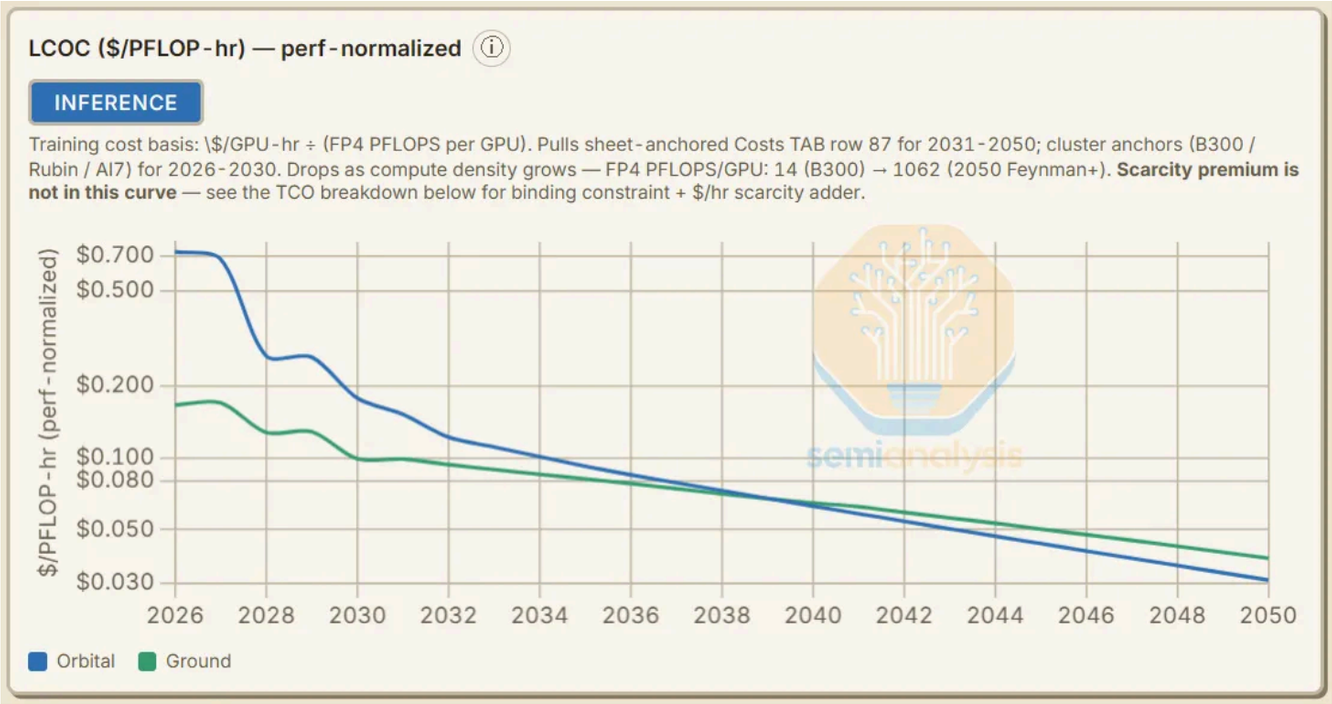
Terrestrial datacenters are clearly more cost effective today, but will that always be the case? When will the effect of technological innovations driving cost efficiencies in space datacenter components as well as the ~80% drop in launch costs from ~$1,400-$1,800/kg for Falcon 9 today to only ~$250/kg for Starship in the future envisioned by SpaceX put costs for space AI datacenters at parity with Earth datacenters?

To answer this question, the AI Space Datacenter TCO Model presents a Levelized Cost of Compute (LCOC) analysis, comparing projections for the cost of compute for space-based and Earth-based datacenters.

In our base case scenario, the cost difference between space and terrestrial datacenters starts at over 4x in 2026, before narrowing to parity in ~2040, with levelized costs of compute in space declining below terrestrial thereafter.

The model and this report will show that **successfully scaling technology and launch costs is a necessary condition for deploying space datacenters**. Once Earth-Space cost differentials narrow meaningfully, then it is the potential shortfall of terrestrial capacity that will drive actual demand rather than further cost differentials.

Another obstacle to overcome stems from chip reliability and servicing. On Earth, 3-6% of GPUs in a cluster annually suffer failures that require human intervention. To ungate space datacenters, we will need to solve this problem either through robotics, greater reliability, over provisioning or a combination of all of the above.



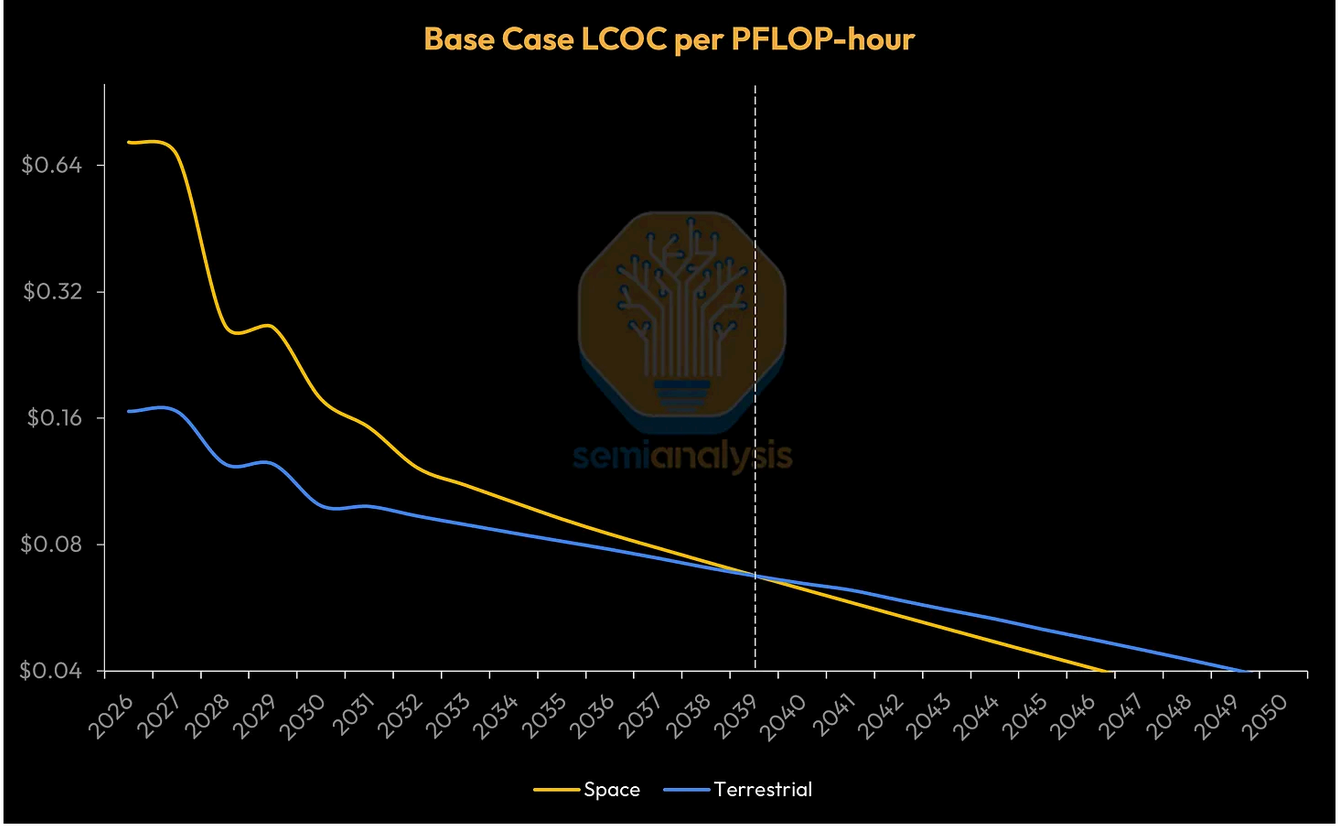
Source: SemiAnalysis AI Space Datacenter TCO Model

In what scenario would the cost differential be narrower or could we even see space datacenters come out as decisively more cost effective? The AI Space Datacenter TCO Model allows users to customize many of the underlying assumptions.

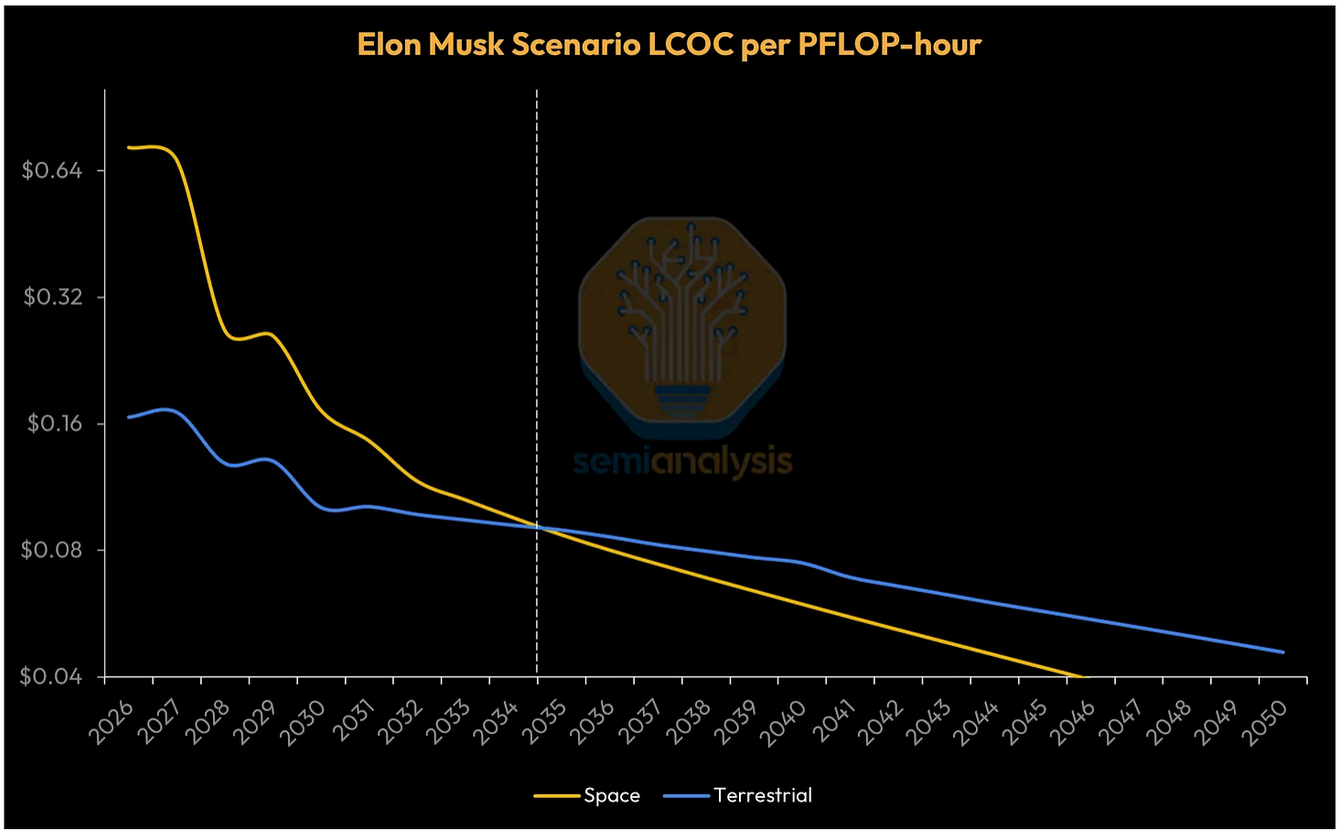
In our base case, we make the following assumptions: (a) critical technology and engineering challenges (namely radiation impact and GPU reliability) are sufficiently solved or mitigated by ~2040, (b) large cost categories (launch, radiators, and solar) also achieve sufficient cost-down scaling, and (c) a generally bullish outlook on overall AI demand and chip production ramp.

The second scenario, the “Elon Musk case”, is where, on top of the base case assumptions, we assume incremental terrestrial datacenter capacity becomes even more difficult to ramp and costly (this premise is supported by rising ground DC input costs already being seen today), which further incentivizes space-based datacenters (i.e., more cost-effective to use orbital compute). This scenario also assumes a successful ramp of an additional ~1,000k wafer starts per month (WSPM) by 2040 added by Musk’s Terafab - in itself no mean feat, though by then we already expect a considerable acceleration in fab capacity expansion. Again - both the base case and the Elon Musk case are upside cases and depart from our industry models which focus instead on confirmed capacity additions.

The base scenario sees Space-Earth cost parity, as measured by Levelized Cost of Compute (LCOC) in $ per PFLOP-hour, by ~2040, while the Elon Musk Case sees near-parity earlier - by the early 2030s.

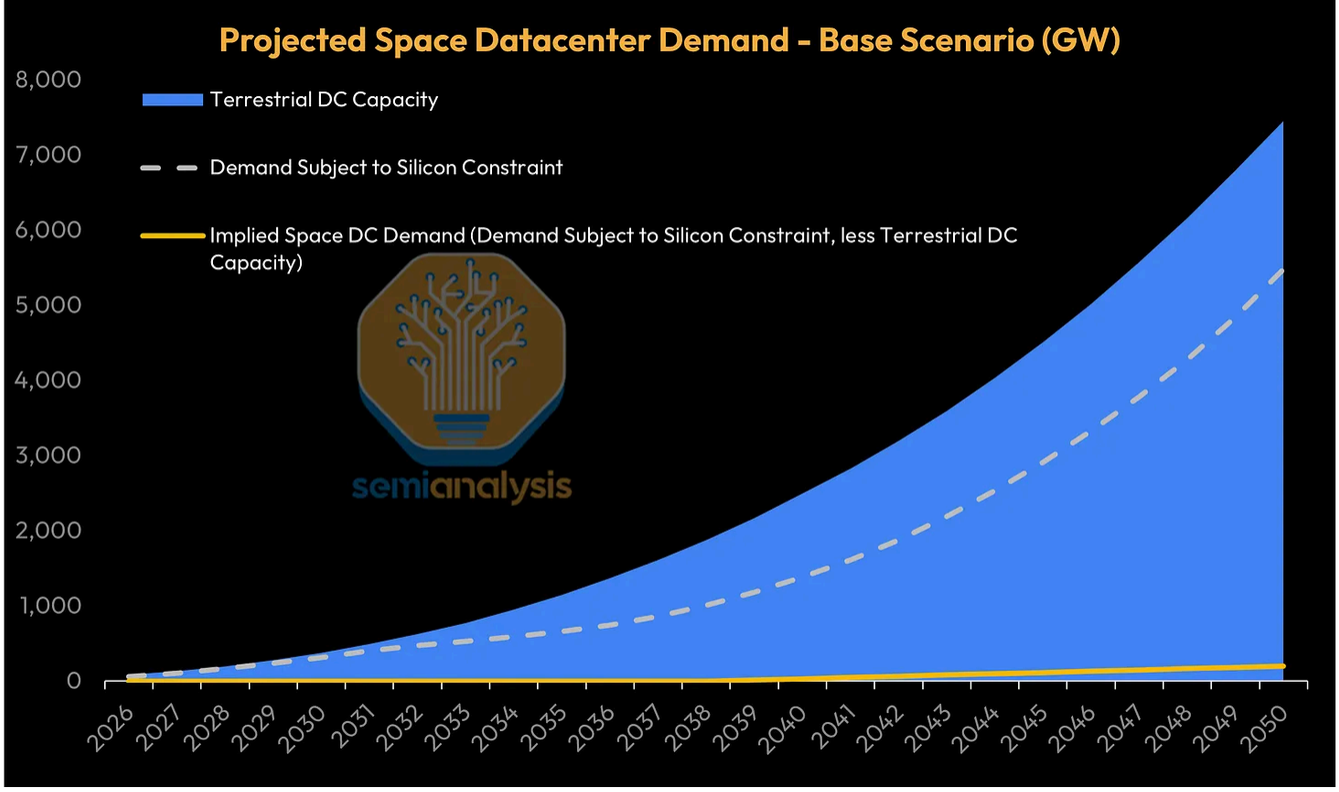


Source: SemiAnalysis AI Space Datacenter TCO Model

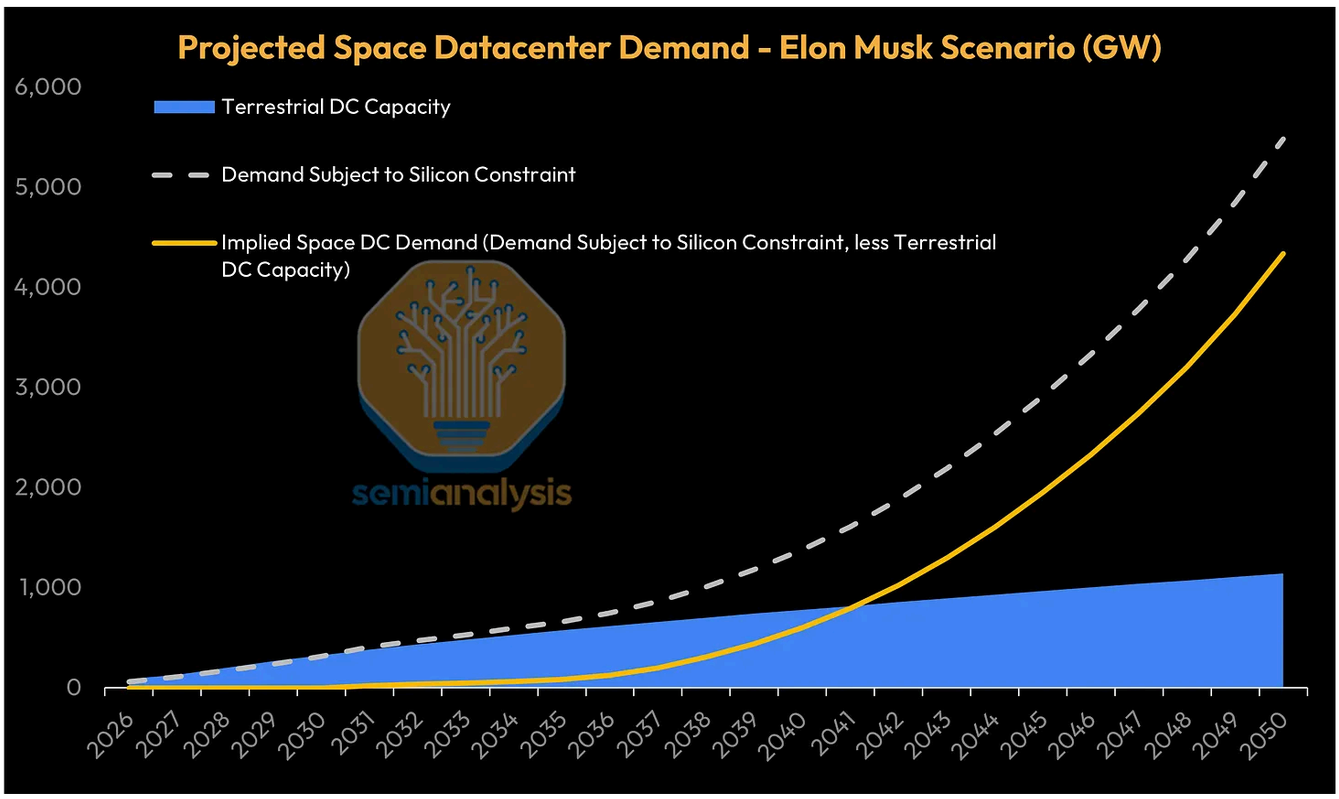


Source: SemiAnalysis AI Space Datacenter TCO Model

Compared to our base case, the Elon Musk case also sees greater and earlier space datacenter demand given less terrestrial datacenter supply to fulfill overall demand (subject to AI silicon constraints).



Source: SemiAnalysis AI Space Datacenter TCO Model



Source: SemiAnalysis AI Space Datacenter TCO Model

There is a lot of ground to cover to support the above conclusion, and in the rest of this report, we will deep dive into the underlying details of this framework and the build up of all the major assumptions underlying this analysis.

This deep dive has four main parts:

Part One is a warm-up round that will debunk a few misleading arguments made regarding space datacenters. This is a great way to introduce readers to important and foundational concepts.

Part Two will focus on discussing the four layers of incremental power supply here on earth after briefly addressing a few of the casual arguments for space datacenters.

Part Three introduces the AI Space Datacenter Total Cost of Ownership (TCO) Model and explains the SemiAnalysis TCO framework that is used in the model to discuss and compare costs for terrestrial and space datacenters.

Part Four explains in detail the build up of a space datacenter with a system by system cost breakdown as well as a discussion of the science behind space datacenter design and implementation.

Taken together, this deep dive will give readers a framework for evaluating future scenarios that may make space datacenters a viable decision and evaluating the arguments for or against space datacenters.

Launching our AI Space Datacenter TCO Model

Today, we are also launching our AI Space Datacenter TCO Model into General Availability. The model provides a first-principles, system-level framework for evaluating orbital compute economics, engineering constraints, and supply-demand dynamics across both terrestrial and space-based infrastructure.

It spans from launch vehicle physics and thermal rejection limits to AI demand curves and GPU-level cost of ownership, enabling users to stress-test when and how compute may transition off Earth. Our model spans 2026 - 2050, with dynamic scenario modeling driven by user-controlled assumptions.

The model will answer the question of if and when space datacenters will become economical and will show the scenarios with respect to end demand and costs on Earth that make Space a viable deployment option.

Please reach out to sales@semianalysis.com to find out more!

Part One: Debunking the Four Casual Arguments for Space Datacenters

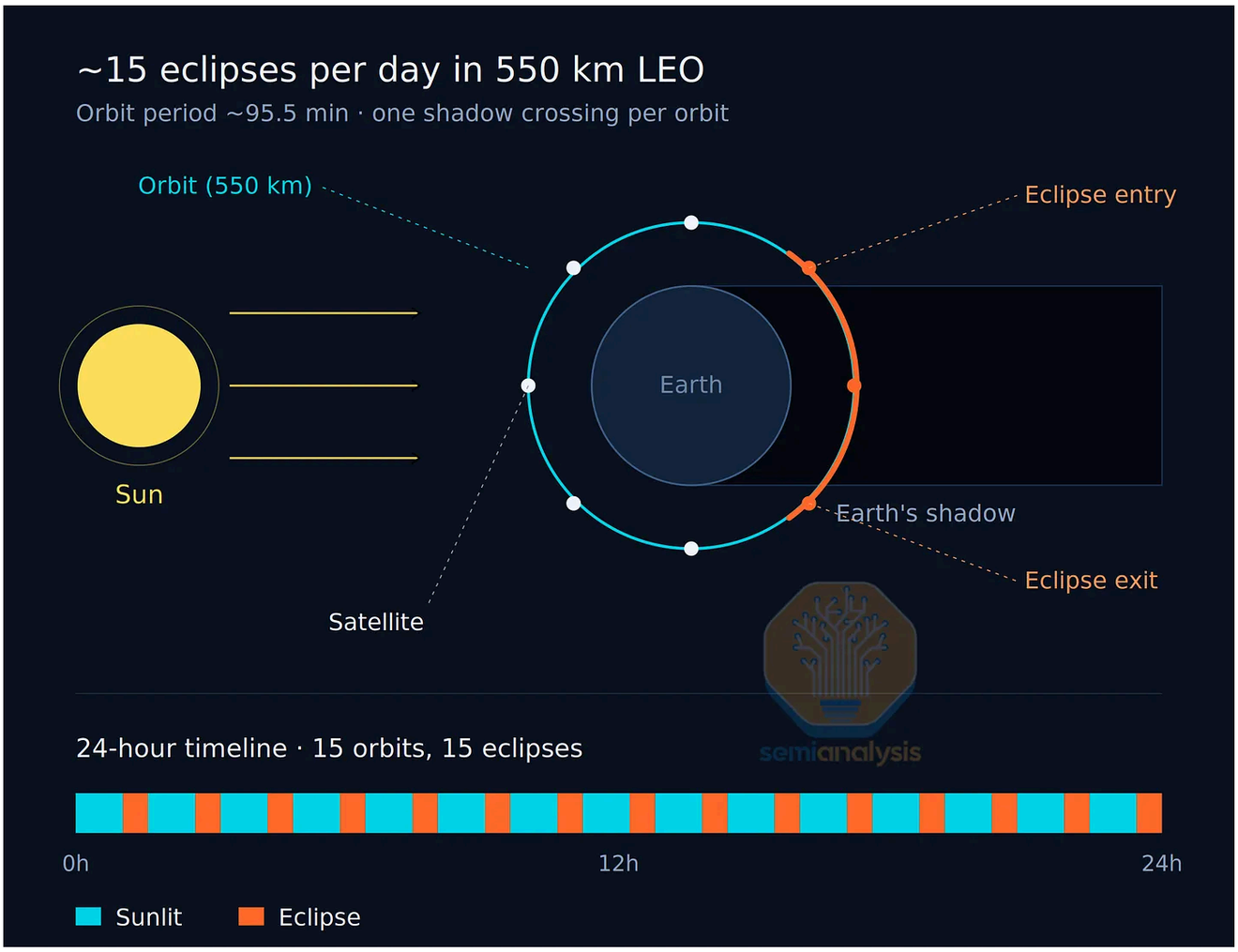
Deploying datacenters in space will be hard, but for different reasons than most think. Let's first debunk the overly simplistic arguments behind these four assertions regarding orbital compute.

Argument #1: You Get 24 Hour Free Solar Energy in Space

Most orbits do not actually achieve 24-hour exposure to the sun. The ISS and majority of the Starlink constellation are in Low Earth Orbit (LEO - within 400-500km of earth) and complete ~15 orbits per day, meaning objects in LEO are receiving sunlight for ~60% of the time. This means that out of 1,361 W/m^2 potential solar irradiance, a space

datacenter in LEO might only capture ~800 W/m^2 in an average 24 hours. The datacenter will also require battery energy storage sufficient to supply 100% of the IT power when the space datacenter is in eclipse, adding hardware cost and complexity.

Instead, a Sun-Synchronous Orbit (SSO) is ideal for Space datacenters. SSO is a retrograde orbit (in the opposite direction of earth's rotation) at a high inclination of >90 degrees that tracks the Earth's terminator and faces the sun most of the day, save for an eclipse of up to 35 minutes per day. The battery energy system required to power the datacenter during eclipse is of far lower capacity than for LEO, but it still entails integrating some power electronics as well as some engineering complexity.



Source: SemiAnalysis AI Space Datacenter TCO Model

Source: SemiAnalysis AI Space Datacenter TCO Model

Argument #2: Cooling is Free

This is perhaps the most misunderstood claim - cooling is decisively not “free”. In fact, it’s quite the opposite. Space is cold, but it is called “Space” for a reason - unlike on Earth, where the atmosphere can conduct heat away from datacenter cooling towers using convection, there is nearly nothing in space to conduct away the heat via convection, and the primary way to remove heat from a datacenter in space is through radiating the heat away. Radiation can remove heat from an object even in the vacuum of space as internal thermal energy leads to the oscillation of charged particles generating photons that remove the heat energy from the object.

Cooling is the largest structural constraint for orbital compute. The International Space Station’s radiator system can only remove 70kW of heat (a quantity about half of what is required to cool a 140kW GB300 NVL72 rack!), requiring a total area of 325 m^2 and a total cost of $340-$500M – more on this later.

Admittedly this system was based on 30 year-old technology with plenty of cost bloat baked in, and costs have improved considerably since then. However, it illustrates the fact that heat removal is one of the core engineering problems to solve when it comes to space datacenters.

Argument #3: You Get The Lowest Latency in Space

Low communications latency is often cited as a key benefit enjoyed by deploying compute in space. Nothing is faster than light in a vacuum. However, the truth is more complicated. A LEO compute satellite completes about 15 orbits per day and is only

over a given ground station for 5-7 minutes per pass. If the satellite is over the ground station nearest to you, then the connection will be strong, but this will only happen for 5-7 minutes per day.

Outside that window, the traffic must go to another satellite in the constellation or go through multiple inter-satellite links to reach a gateway that is closer to the end user. If there is a satellite passing over the Indian Ocean that is serving a US Customer, those hops through multiple Inter Satellite Links (ISLs) can accumulate 30-80ms of one-way delay. This problem worsens when you switch to optical ground links instead of RF as there are a limited number of ground stations at any given time that do not suffer from some degree of atmospheric interference. This necessitates the deployment of many ground stations spread throughout the world. Communicating from these dispersed ground stations to the end user is yet another source of added latency.

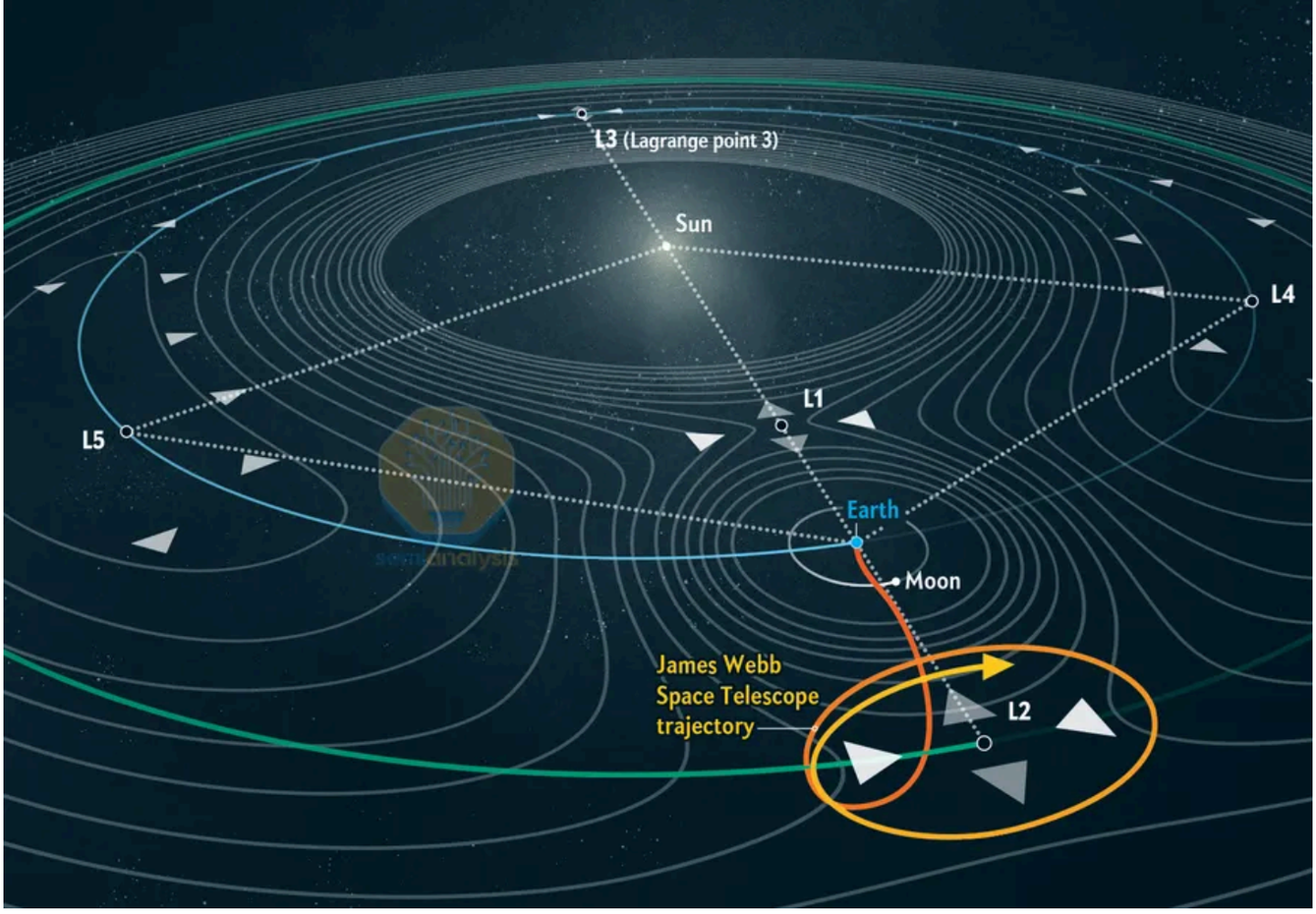
Argument #4: No Need for Permitting in Space

Permitting is more than just a headache when building a datacenter on Earth. In the US, projects face transmission cost inflation and study-related uncertainty, on top of the 5-7 year waits to connect to the grid. As we have outlined in the latest major expansion of our Energy Model, there has been substantial growth in onsite gas and behind-the-meter (BTM) projects as a workaround to waiting in the interconnection queue, but air permitting remains a significant hurdle. Wait times for turbines are also long, with the main OEMs, such as GE Vernova and Siemens Energy, having limited capacity to serve projects planning to come online before the end of the decade.

When it comes to space, one wrinkle is that SSO is a constrained subset of LEO, rather than a separate orbital regime. LEO spans 400-2,000 km across many inclinations, with carrying-capacity estimates ranging from ~100,000 to over a 1M satellites depending on separation and collision-tolerance assumptions. SSO, by contrast, requires a specific altitude-inclination relationship (roughly 97–99° inclination at 500-1,000 km) so that the orbital plane precesses at the rate needed to maintain a fixed local solar time.

Most operational SSO traffic concentrates in a 600–800 km sweet spot. Dawn-dusk SSO, the specific subset that rides the terminator and is relevant to orbital compute, is even narrower. It represents a single local-time slot within an already constrained orbit class, and is materially smaller in usable capacity than LEO.

Other than dawn-dusk SSO, there are other “orbits” that are exposed to the sun 24/7 - for example the Sun-Earth Lagrange Point L1. But here latency is a clear deal-breaker - it takes light ~10 seconds to travel the 3 million km round trip between Earth and L1.



Source: Scientific American

Part Two: Terrestrial Datacenter Constraints and the Four Layers of Incremental Power Supply on Earth

The Four-Layer Framework: How Constrained is the Terrestrial System?

In a recent appearance on the Dwarkesh Podcast, Elon Musk promoted the idea of orbital compute, arguing that Space would shortly be the most compelling place to deploy AI. Musk's argument is not just that space-bound chips may matter eventually, but that terrestrial power, turbines, and permitting will hit a wall fast enough that space could become the cheapest place to run AI within roughly three years. There are warnings from the turbine manufacturers, and queues to connect to US grids that could last years, so you can see why he thinks this.

Musk and SpaceX clearly believe in this and have put their money where their mouth is, with corporate commitments rewarding Musk with up to 302M Class B SpaceX shares (between ~$30B and ~$60B USD assuming a $100-$200 IPO price range) for delivering 100 TW per year of non-Earth based datacenters – though we will size the analytical realism of that milestone later in this report. On top of that, the recent launch of the Terafab initiative acknowledges one other additional constraint: you can't build a datacenter anywhere if you don't have the chips to fill it with.

To stress-test Musk's power argument on its own terms, it's worth asking: **even if chip supply were unconstrained, would terrestrial power suffer constraints so severe as to force AI compute into orbit?**

Borrowing the Peak Oil Framework

The "Peak Oil Theory" framework is helpful in answering this question. In the 1970s, this theory that global oil production was nearing a hard geological ceiling, and that oil would soon become impossible to produce and use. Instead, new supply sources became economical as conventional supply tightened and oil prices increased, boosting the incentive to drill harder. Supply moved up the cost curve and became more infrastructure-intensive, and in more recent decades, the technology improved to access more oil supply, stabilizing prices.

That framework can be applied to datacenter power, where the sources of power supply are varied and the barriers to entry are a lot lower than the much more consolidated chip manufacturing space, which has plenty of near-monopoly producers across the supply chain.

Let's borrow from the Peak Oil Theory as we walk through the terrestrial power stack layer by layer. As we move through the latter half of this decade, power access is set to get scarcer and trickier to execute on even when available - the end result is that power becomes more expensive. Cost escalates as we tap into more difficult to access sources of power.

All of these layers will have to become exhausted before space becomes an economically viable alternative, or even a preferred option. The fifth layer is a "universal" constraint on all chip deployment, whether deployed on Earth or in Space.

Terrestrial Power Stack			
Layer	**Description**	**Capacity**	**Cost Range**
Layer One	Grid-connected supply: Hyperscaler self-build, colocation, Neocloud	~81 GW currently to ~214 GW in 2030	$12–15M/MW
Layer Two	Bitcoin-to-AI conversions: Converted bitcoin miners and powered land	~2 GW currently to ~10 GW in 2030	$10–15M/MW
Layer Three	Behind-the-meter generation: Gas, fuel cells, nuclear	~3 GW currently to ~97 GW in 2030	$15–20M/MW
Layer Four	Industrial production expansion: More copper, transformers, labour	~3 GW currently to ~18 GW in 2030	$20+M/MW
Layer Five	Semiconductor ceiling: Chips, not power, are the binding constraint	N/A	N/A

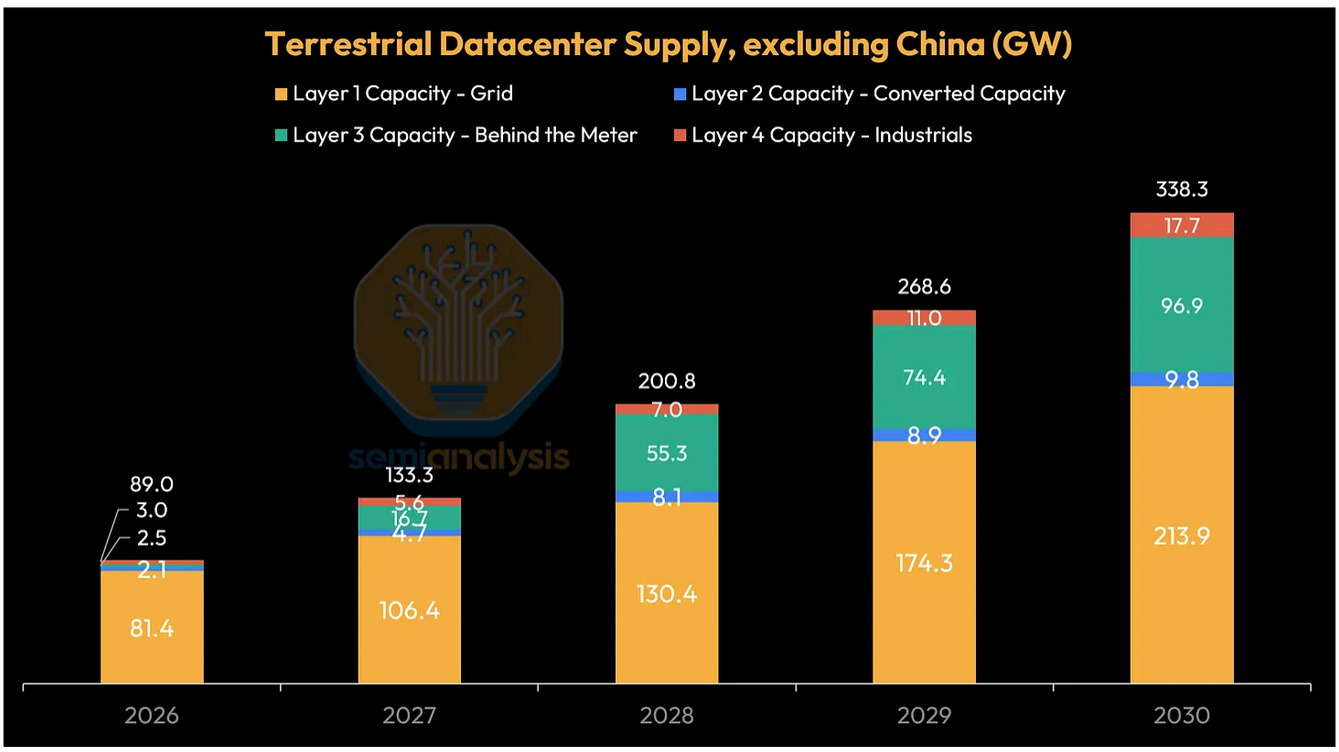
Source: SemiAnalysis AI Space Datacenter TCO Model, SemiAnalysis Energy Model

While datacenter capacity was the bottleneck in the past few years, the overriding constraint today has now moved up the stack into semiconductor manufacturing —

specifically advanced-node logic capacity at TSMC, HBM production at SK Hynix, Samsung and Micron, and DRAM output across the industry.

Moving giga- or even terawatts of chip capacity into space is a non-starter if we cannot even manufacture the chips to put up there.

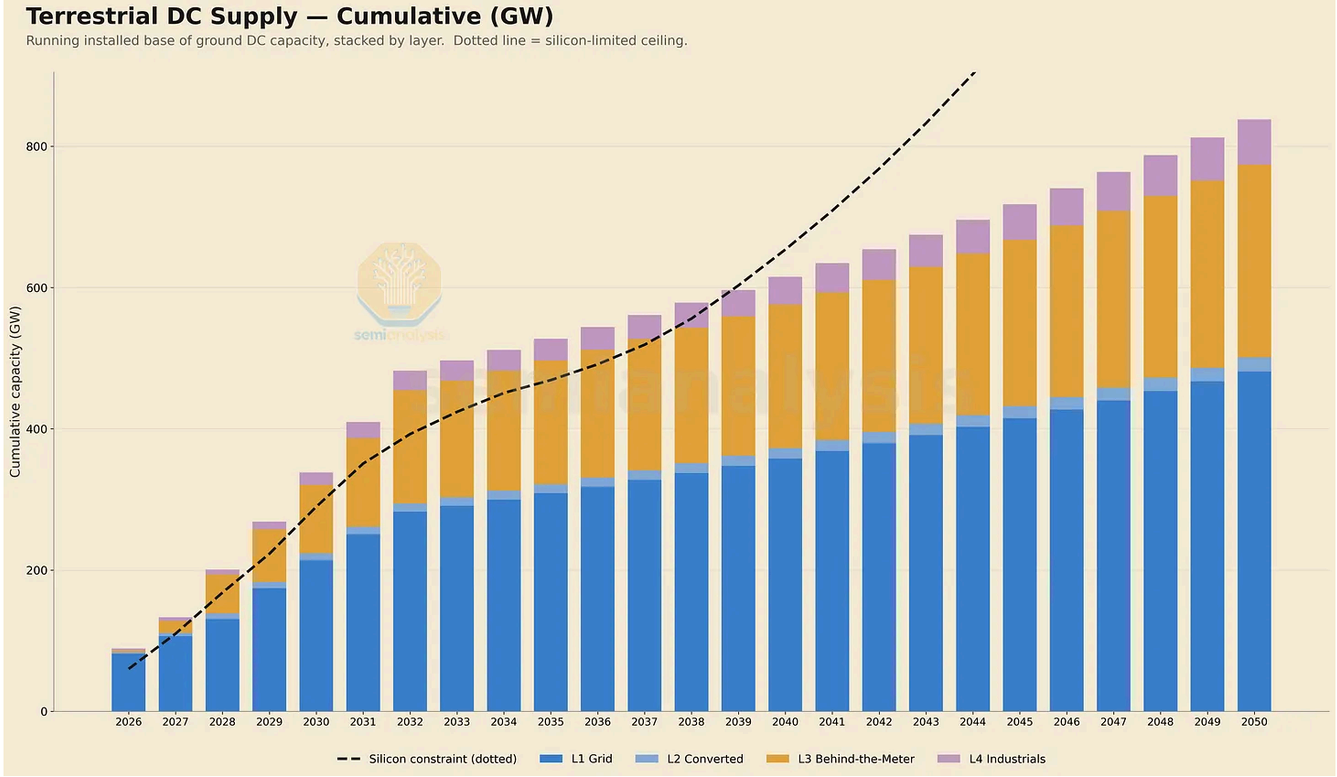
As chip manufacturing constraints begin to limit the pace of AI accelerator production, we will now counterintuitively have sufficient datacenter supply in 2027 to cater to chip demand. Our SemiAnalysis AI Datacenter Model models confirmed facility and land bank pipelines, and in the AI Space Datacenter Model, we take a more optimistic capacity view by including potential capacity sources that have yet to be enumerated. Adding layers one through four show that total tracked global datacenter capacity (excluding China) could rise from 89 GW in 2026 to up to 338 GW in 2030 - nearly quadrupling overall capacity over four years. However, this capacity is delivered by tapping layers two to four to allow for more datacenter capacity, and accordingly we see converted capacity, behind-the-meter (BTM) generation, and industrial production expansion kicking in.



Source: SemiAnalysis AI Datacenter Model, SemiAnalysis AI Space Datacenter TCO Model

Current projected builds and allocated capacity implies tens of gigawatts of headroom in the next few years vs planned accelerator deployment. Unfortunately, **Accelerator demand will be bottlenecked by chip production capacity for the next few years, and this is a problem that Space Datacenters cannot solve.**

The below table from the SemiAnalysis AI Space Datacenter TCO Model illustrates how silicon capacity holds back AI chip deployment even when more datacenter capacity can be stood up.



Source: SemiAnalysis AI Space Datacenter TCO Model

Notwithstanding the above, let's step through each layer of supply and discuss the constraints involved for each layer in detail.

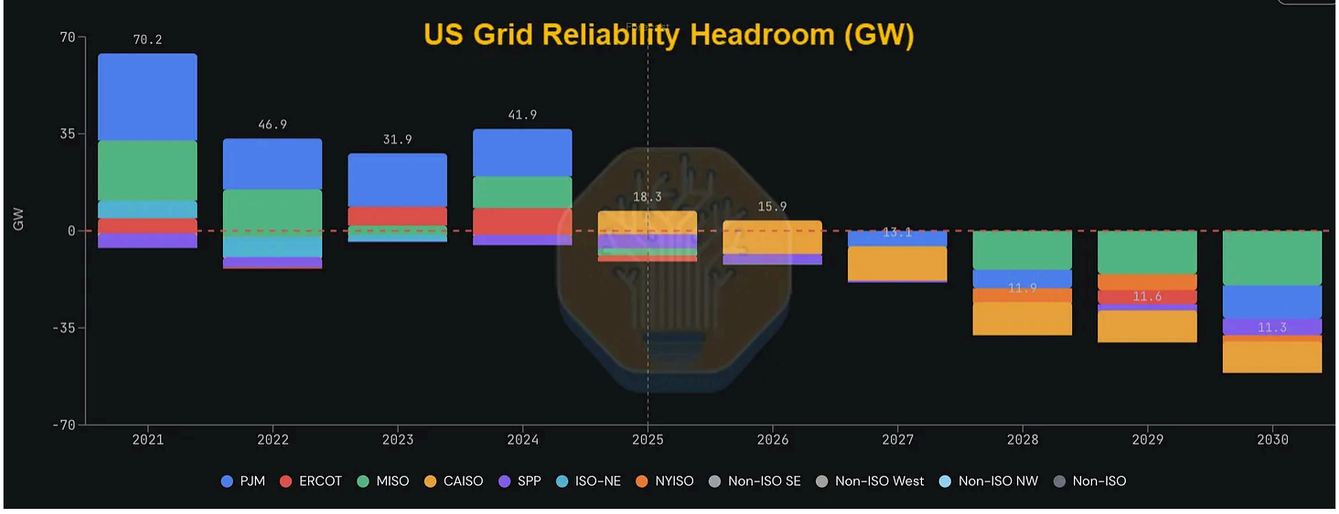
Layer One: Grid-Connected Supply

Grid-connected power is still the cheapest on paper, with infrastructure costs in the $12–15M/MW range. But in most US markets, the real cost is waiting in queue for power interconnection from the grid. North Virginia's PJM interconnection timelines now run to roughly seven years in practice and that timeline is not workable for hyperscalers that can barely keep up with end demand.

Of course there is an element of overbooking interconnect capacity, but even discounting for speculative filings, the approved and under-construction pipeline is large enough that physical transmission and substation constraints in the major build markets constrain the pace of power addition for the next few years. More on this can be found in our article on the Electric Reliability Council of Texas, i.e. ERCOT's new batch study process.

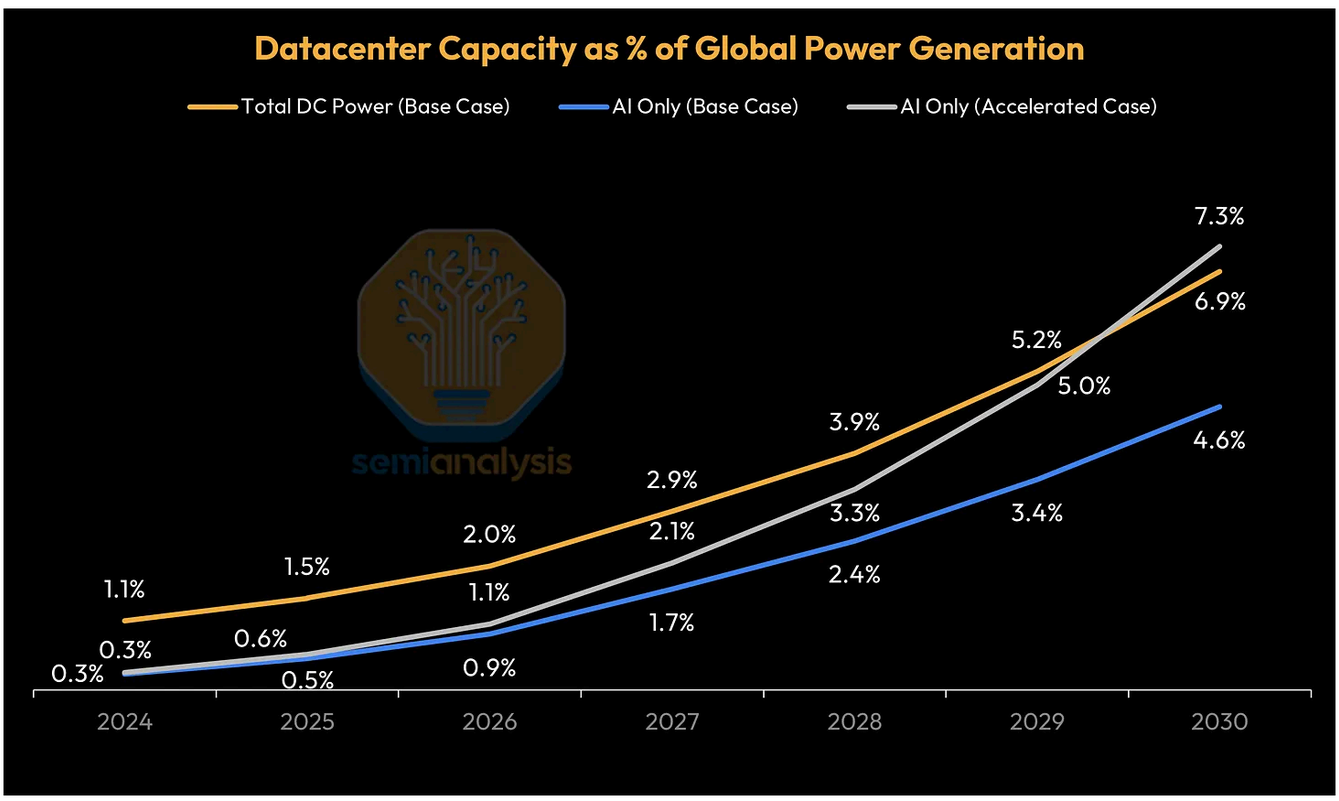
Across US ISO territories, gross positive grid reliability headroom (defined as the buffer of effective supply above peak demand plus required reserve margins, measured on SemiAnalysis's penetration-adjusted ELCC methodology) stood at roughly 18.3 GW in 2025, down from 70.2 GW in 2021. In practical terms, grid reliability headroom is a metric that is linked to how much capacity can be connected to the grid at all.

That buffer narrows to just 15.9 GW in 2026 before net headroom turns negative in 2027, reaching an aggregate deficit of approximately 40 GW by 2030. Negative headroom means the grid is oversubscribed on a planning basis - more load is projected than accredited supply can reliably serve under each ISO's reserve margin requirements.



Source: SemiAnalysis Energy Model

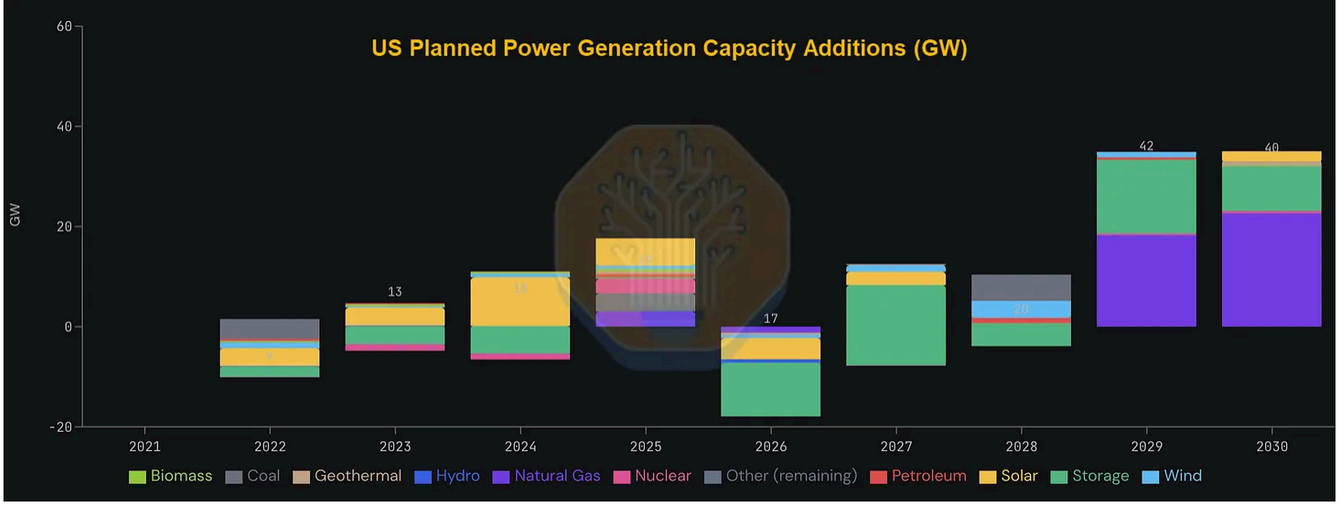
From a global energy perspective, datacenter electricity consumption reached nearly 340 TWh in 2024 - roughly 1.1% of global power output. AI-specific demand accounts for around 0.3% of global generation today, crosses 1.7% around 2027 in our base case, and reaches just under 5% by 2030. Even in an accelerated scenario it stays around 7% by 2030, with that ceiling representing roughly 380 GW of continuous power demand.



Source: SemiAnalysis Energy Model

Despite the aforementioned tightness, help is on the way.

US planned power generation capacity additions are stepping up sharply in the coming years. Raw nameplate overstates what is actually usable, so we use Effective Load Carrying Capability (ELCC) to measure how reliably each source can meet peak demand. This adjusts for intermittent solar and wind power sources which contribute far less to the grid's firm capacity than their headline numbers suggest.



Source: SemiAnalysis Energy Model

This is a picture of progress. The US grid is adding capacity, solar is coming in faster than any prior generation technology, and international buildout continues to absorb demand across EMEA and APAC.

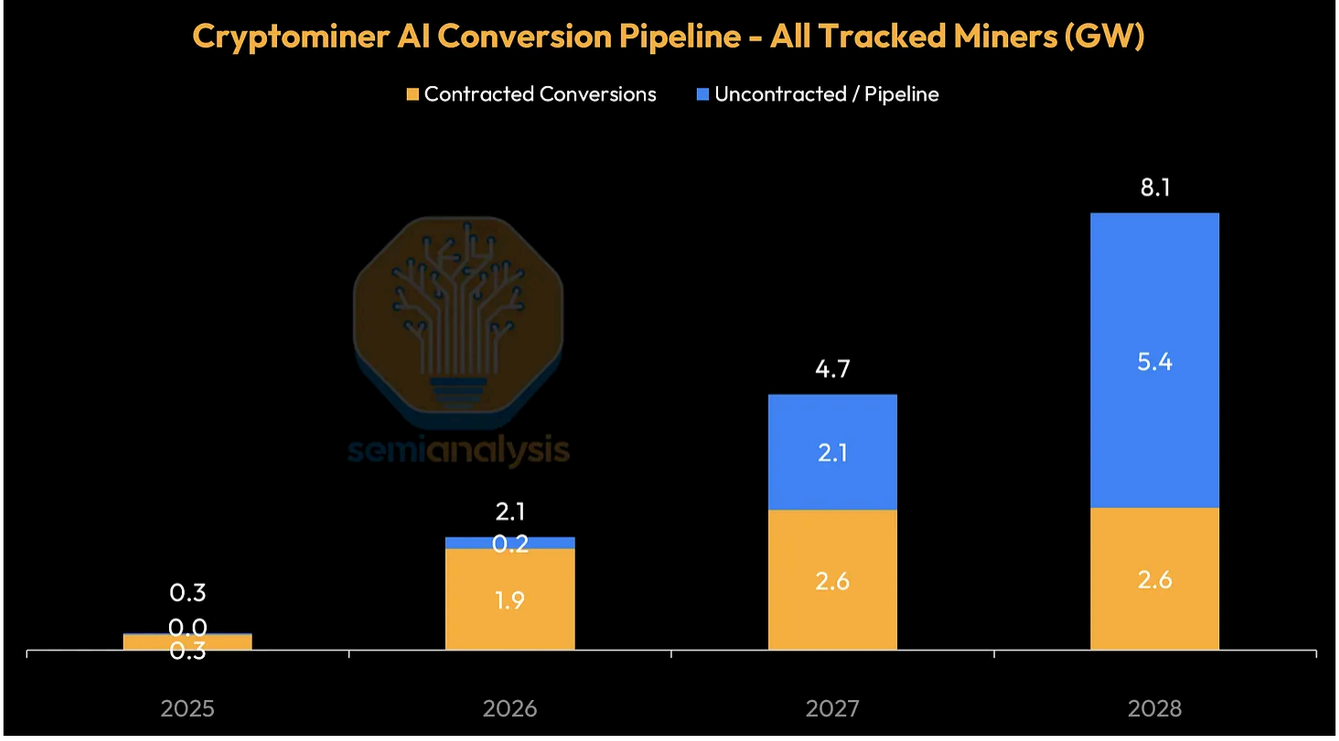
The ~106 GW of global datacenter capacity by 2027 generation additions and grid-connected supply can keep pace with accelerator demand. Beyond that, the grid alone cannot carry the load and we are pushed into the next layer of capacity addition.

Layer Two: Converted Capacity and Powered Land

When looking beyond the grid - the first place to look is converting existing facilities with power into AI Datacenters. Cryptominer conversions are the clearest example of this second supply layer. Core Scientific, IREN, Cipher Mining, Applied Digital, and TeraWulf together point to roughly 2 GW of tracked contracted conversions by the end of 2026 and roughly 5 GW by the end of 2027.

These are sites with existing grid connections, permitted substations, and in some cases usable cooling infrastructure. Although the quality of the existing technology on-site can vary, leading these sites to be similar or even lower in cost to grid-connected supply at around $10–15M/MW. Firms like Fermi Energy and Cloverleaf are bringing large, grid-connected sites to market by taking interconnection risk themselves rather than leaving it to hyperscalers. Oracle has committed at least $3.65B to support 1.4 GW of capacity at the Related/Oracle/DTE Electric campus in Michigan. That would have sounded extreme not long ago, but looks rational now that 1 GW of AI capacity can support on the order of $12–13B of annual revenue.

In aggregate, converted sites and powered land can add 8-10 GW of near-term supply, with Cryptominer conversions supplying 8GW cumulatively by 2028. This near-term supply can act as a relief valve for a few years before the repurposable inventory is mostly absorbed, and the industry turns to the next layer: off-grid power supply.



Source: SemiAnalysis AI Datacenter Model

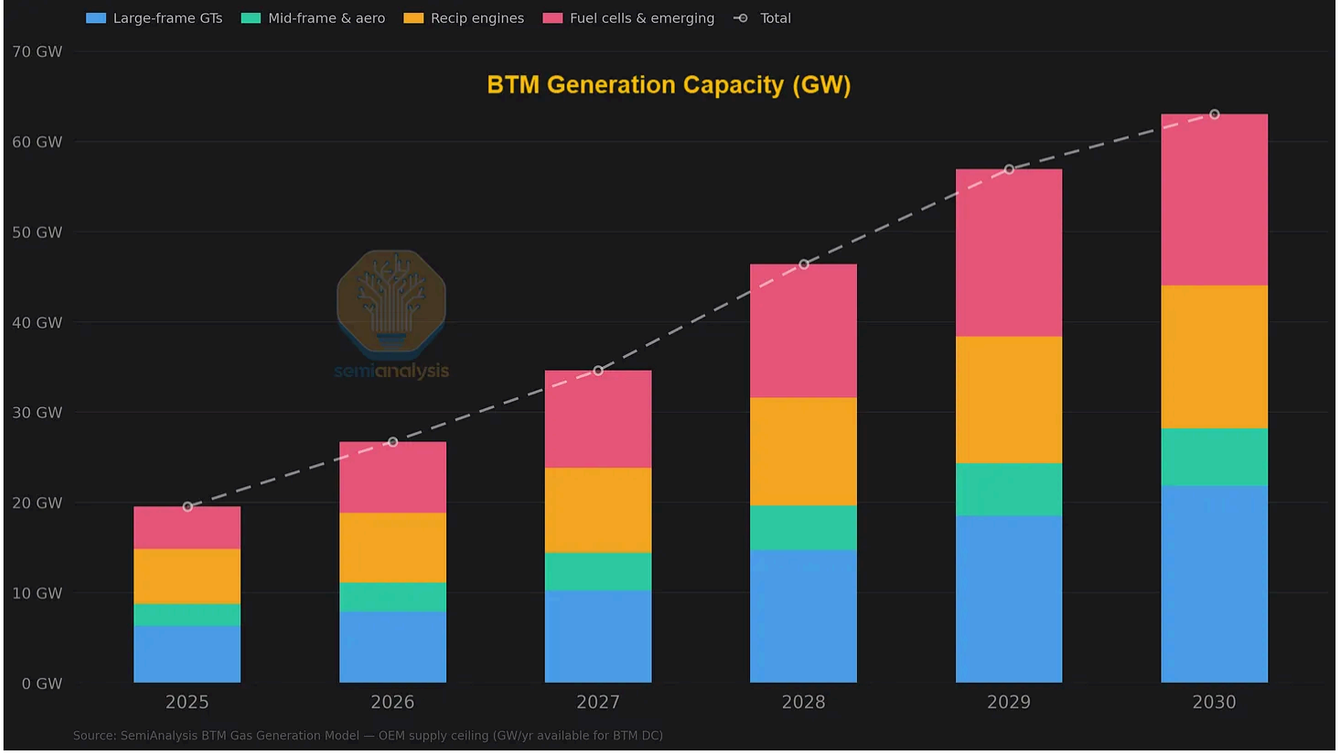
Layer Three: Behind-the-Meter (BTM) Generation

BTM generation was once considered a last resort, but when major AI cloud contracts imply annual revenue in the range of roughly $12–13M per MW of contracted critical IT load, getting 200 MW online six months early carries an NPV of roughly $400M-$500M. So any additional capex from bringing your own generation (BYOG) and skipping the queue, can result in quick reward. The all-in BTM costs can still fall in roughly the $110–170/MWh range depending on technology, not dramatically different to grid power that can already clear around $150/MWh in major US markets.

BTM annual additions in critical IT power terms — predominantly US-based — are set to be the primary source of power for half of new AI datacenter power capacity additions by 2028 — up from fewer than 7% of capacity added in 2025. Cumulative confirmed BTM critical IT capacity is set to reach 26 GW by the end of 2030, but this number will likely be much higher still from unannounced projects, with OEMs stating the majority of their orderbooks and enquiries come from datacenter projects. Small modular reactors for BTM operations could come into play, with perhaps 1–3 GW of supply post-2030.

The system-wide DC-addressable ceiling across all BTM generation categories runs into the tens of gigawatts per year by 2027 at a cost of $15-20M/MW — spread across six or seven distinct supply chains that don't all tighten at once.

The BTM market is growing incredibly quickly and existing OEMs and new entrants alike are flooding into the sector. We will soon be launching an expansion to the SemiAnalysis Energy Model, tracking more than 30 OEMs, forecasting quarter by quarter manufacturing capacity, installations, and availability.



Source: SemiAnalysis Energy Model

Layer Four: The Production Constraint in Industrials

The headroom for more compute on Earth is there from a power and shell perspective once we pull the levers of converted capacity and BTM generation, but the next layer, Layer Four: Industrial Production, consists of supply created by mustering capital for building additional manufacturing to enable more provision of all the prior layers.

Large power transformers already have some of the longest lead times in the electrical stack, this is because building transformers depends on a small number of grain-oriented electrical steel (GOES) producers globally. Transformers are generally required in both grid-fed and BTM scenarios alike. Copper is another restriction, given how broadly it sits across transformers, turbines, cables, busbars, and cooling equipment. Copper is difficult to mine, and has seen a nearly 20% price jump over the past year. That said, the increase in prices further downstream to the transformer manufacturers, are mostly demand-signal driven, and not reflective of actual supply. And the transition to HVDC and optics in networking are increasingly being considered as antidotes to this issue.

Labor for construction and operations is getting more costly, but modularization and digitalization of datacenters has already been implemented in buildout, and can reduce on-site labor needs by more than 50%. Still, a substantial number of skilled man-hours remains, and this would compound substantially once we head into the multiple hundreds of GWs of compute territory.

Tapping into this layer would push costs per MW beyond the $20M/MW range - how far beyond will depend on how many tens or even hundreds of gigawatts of capacity but how costs escalate once we are in this fourth layer will determine whether space datacenters make economic sense.

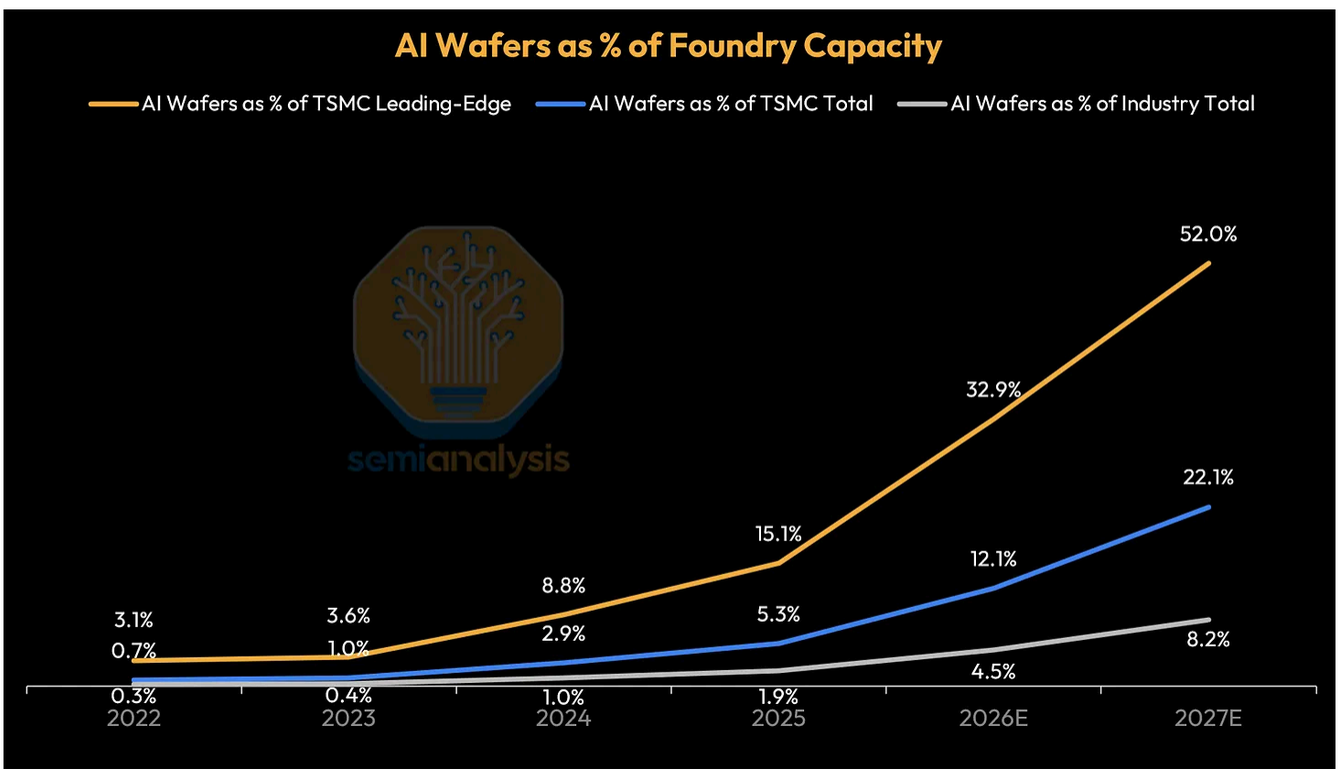
There are a wide variety of outcomes cost-wise. Users of our new AI Space Datacenter TCO Model can adjust assumptions around base Layer Four costs and the price elasticity of supply to analyze various scenarios.

Layer Five: The Semiconductor Ceiling

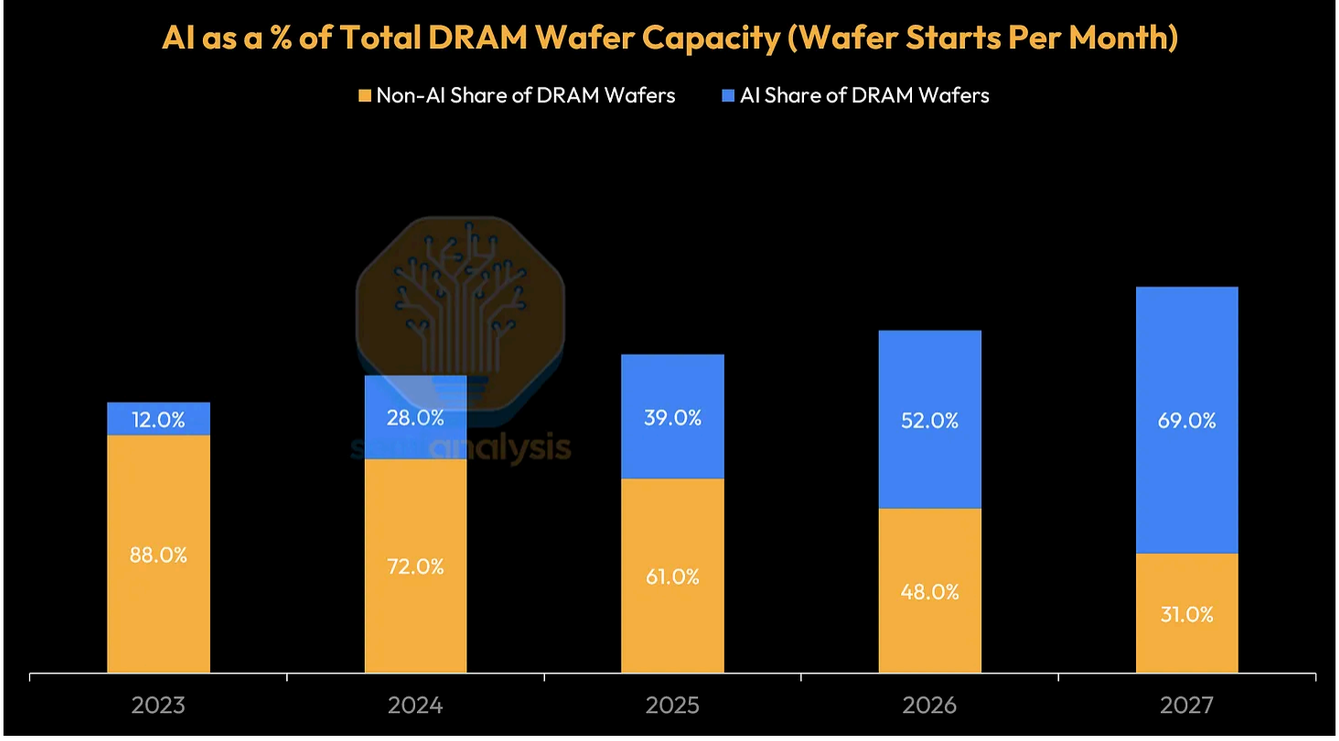
In the podcast with Dwarkesh, Musk does not name the hardware that would be best suited to his stellar ambitions. But does this even matter? As we have established, ramping up production for any chip is also the biggest constraint to mass cluster buildout on Earth. Given trends in Semiconductor production tightness - we will hit a silicon capacity constraint well before the other layers.

AI-related demand is modeled to consume just under 60% of TSMC's N3 output in 2026 and approximately 86% in 2027, nearly squeezing out remaining smartphone and CPU demand. Additional fab area must first be built before chips can be projected into orbit.

Memory faces the same problem from another angle: incremental DRAM capacity is increasingly absorbed by HBM, which consumes roughly three times the wafer capacity of commodity DRAM on a bit per wafer basis basis. Our recent report on silicon and memory shortages goes deeper into this. On the DRAM side, AI-related demand is projected to consume roughly 70% of total DRAM wafer capacity by 2027, up from 12% in 2023 — a near six-fold increase in four years.



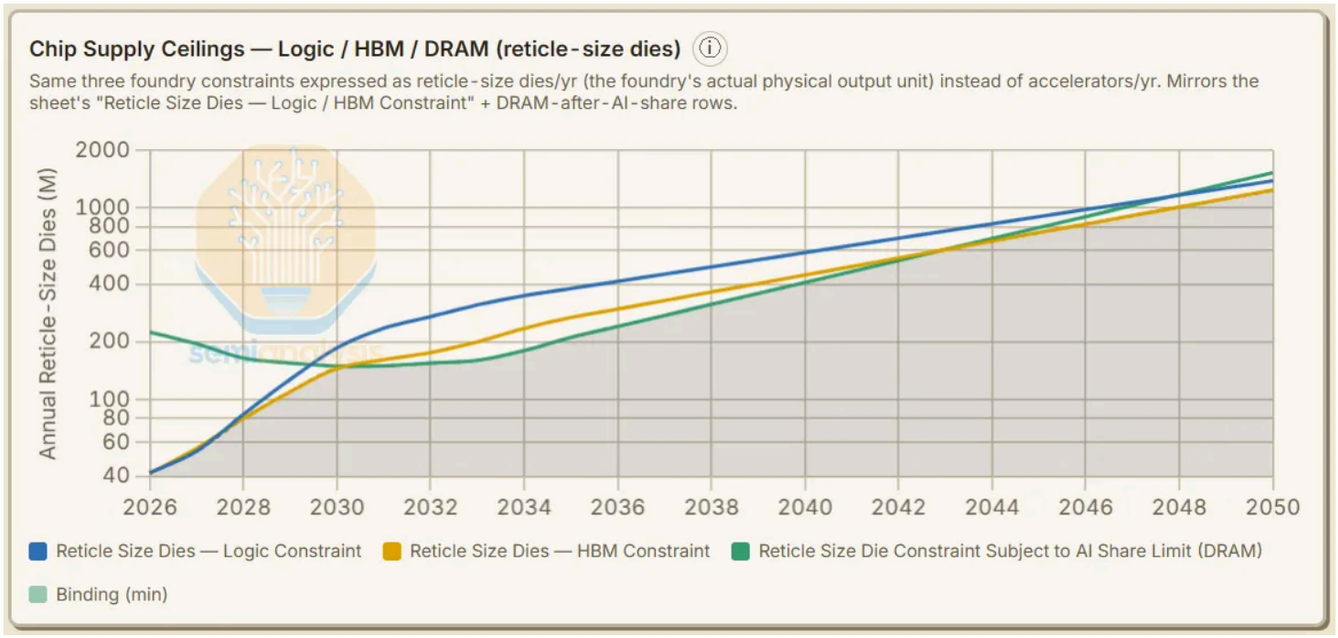
Source: SemiAnalysis Foundry Model



Source: SemiAnalysis Memory Model

The structural reason this ceiling is harder to move than the power stack is cleanroom physics. Adding advanced fab capacity requires building cleanrooms first, then installing tools, then qualifying processes — a sequence that is hard to meaningfully accelerate in the near term regardless of capital availability. The point where that changes looks more like 2032–2034 than 2027–2029, as we will look into further.

In our AI Space Datacenter Model base case - logic and memory capacity additions move in lockstep, though memory tends to be the binding constraint for much of our forecast horizon.



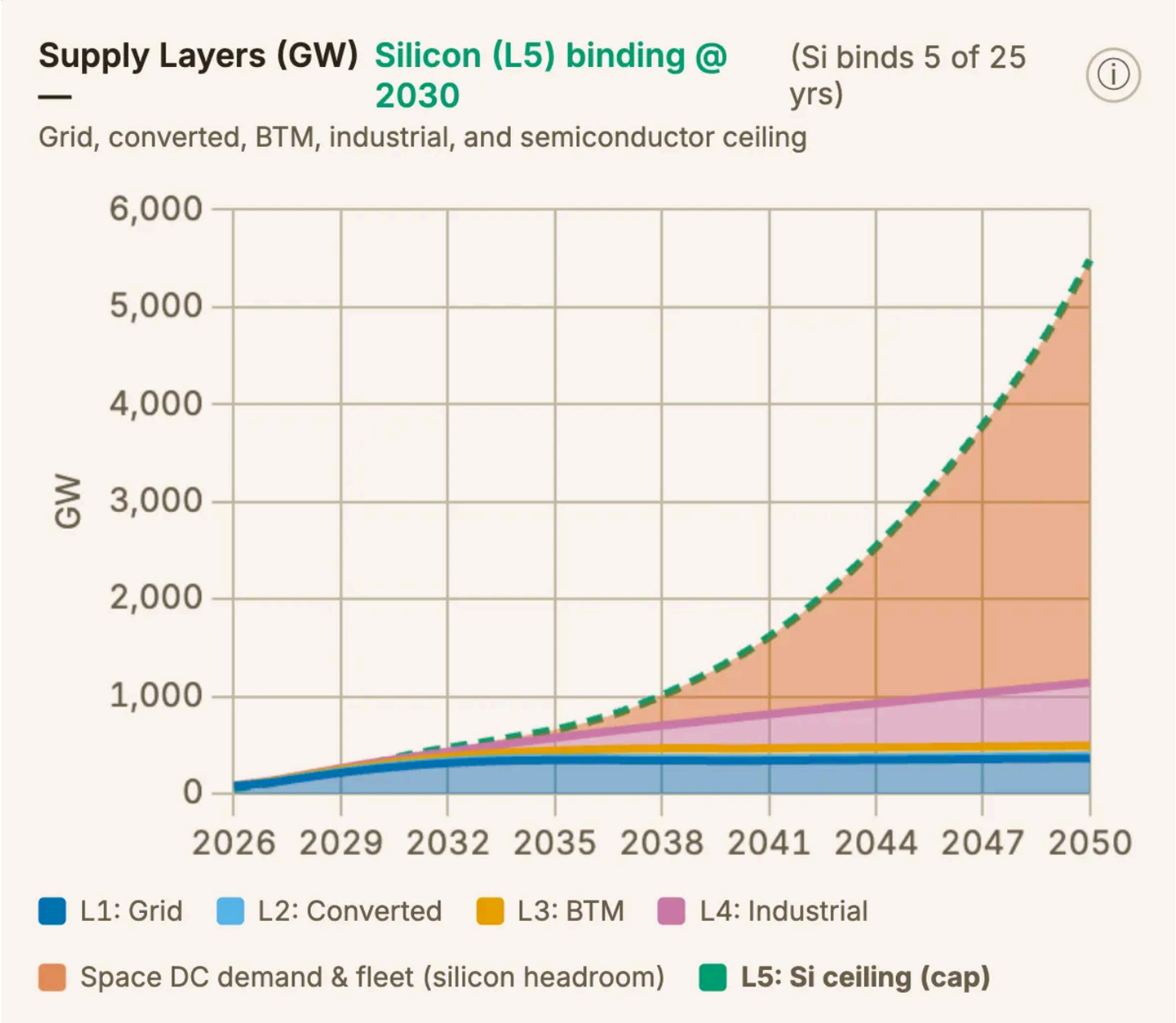
Source: SemiAnalysis AI Space Datacenter TCO Model

Our base case assumes that there are considerable resources mustered to add semiconductor manufacturing capacity - this is clearly not a conservative assumption and there are plenty of bottlenecks that could derail this base case. Yet - even with these loosened assumptions, the silicon constraint is binding in the long-term even

though we forecast it supporting hundreds of GW of AI capacity additions in our base case.

Source: SemiAnalysis AI Space Datacenter TCO Model

In the Elon Musk Scenario - we keep the same robust chip manufacturing capacity expansions, and with much lower terrestrial datacenter capacity, the AI industry must shift deployments to space.



Source: SemiAnalysis AI Space Datacenter TCO Model

Getting even close to fulfilling 800 GW of AI demand will require the entire global EUV fleet dedicated to AI with nothing left for phones, PCs, or anything else. Some semiconductor relief could materialize in the 2032–2034 window, as TSMC's Arizona and Japan capacity comes fully online and memory fabs respond to sustained HBM demand with dedicated capacity additions.

Musk's Terafab Initiative

During the March 2026 launch, Elon Musk framed Terafab as a 1 terawatt per year compute factory. Tesla, SpaceX, and xAI would build it together in Austin on a $20-25B budget, starting at 100K wafer starts per month and climbing toward 1M wafer starts per month (WSPM), roughly 70% of TSMC's entire global output today.

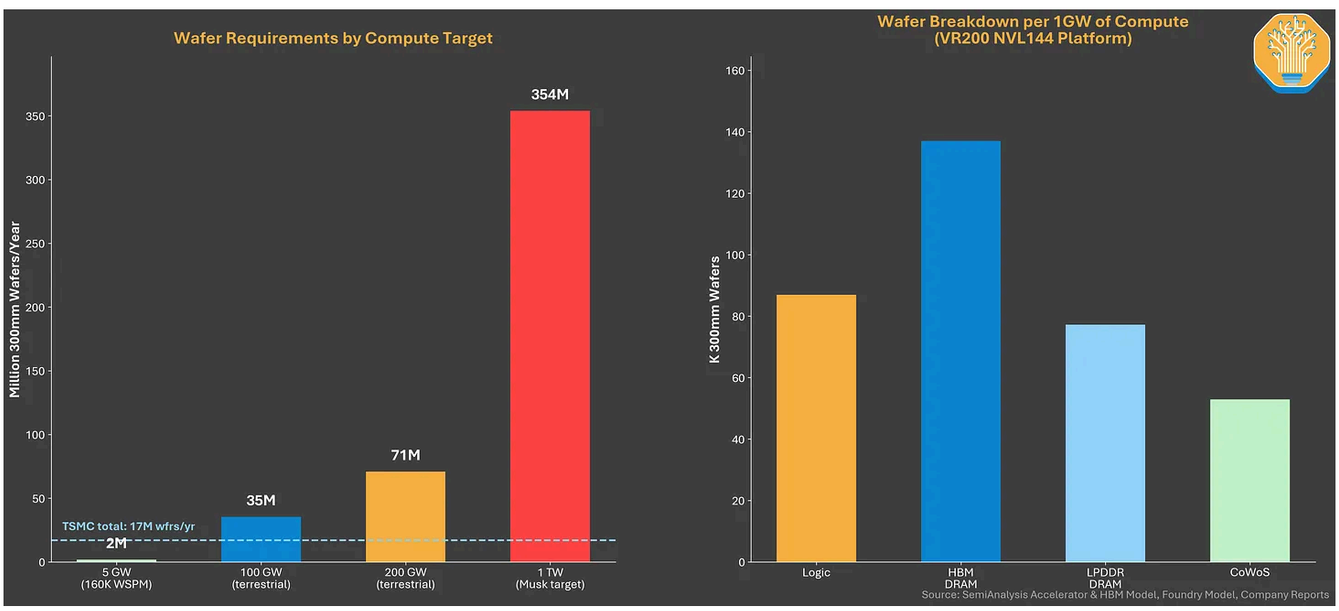
The plant sits on 100 million square feet of floor space, which Musk compared to 10x Giga Texas or 15 Pentagons, across thousands of acres, and draws more than 10 gigawatts. Scope covers logic, memory, mask shop, advanced packaging and testing. Compute allocation splits 80% space and 20% terrestrial, roughly 800 GW for orbital datacenters and 100-200 GW for terrestrial inference - this aligns with SpaceX's recent S-1 disclosures to have one chip variant optimized for terrestrial edge and inference (e.g. Tesla's Optimus robots, vehicles), and another chip variant for orbital compute.

First wafers are claimed to be out as soon as next year in 2027, with mass production in 2028.

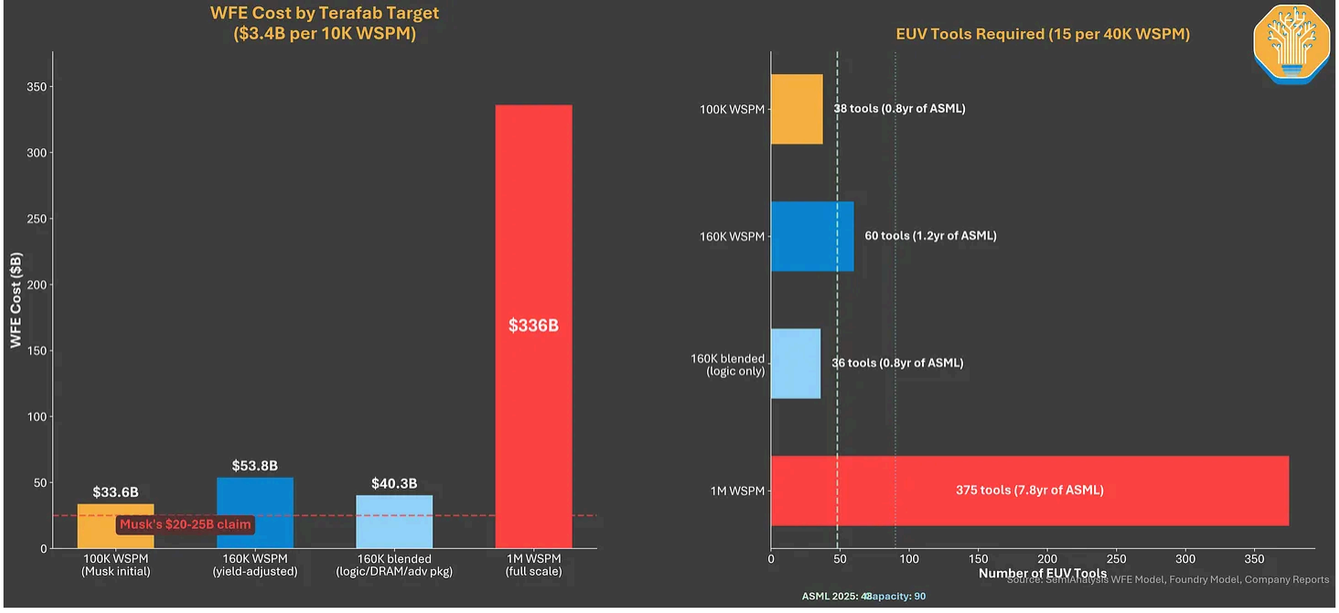
We are not one to bet against Musk. He built reusable rockets, scaled EV production to nearly 2M cars a year, and designed custom inference chips that compete with the best in the industry. Even a fraction of Terafab is a meaningful success. The numbers still matter.

Start with capacity. Our Foundry Model puts global 300mm foundry capacity at 4M+ WSPM in 2025. Terafab at the 100K entry ramps in as 2.5% of the entire world of foundry wafer starts. Terafab at 1M full scale is 24% of global foundry capacity, or 68% of TSMC alone. Nobody has added that much in a decade outside of TSMC itself, and TSMC took three decades to get there.

Our Accelerator and HBM Model puts 1GW of deployed compute at 354K wafer starts across logic, memory, and packaging, with memory over 60% of the total. Wafer value per GW runs near $3B. If "1 terawatt" means 1TW of simultaneously deployed compute, that is 354M wafer starts a year, or 21x TSMC's entire global output. If it means cumulative installed base over 15 to 20 years, the math works but only barely. Our best guess is that "terawatt" plays the same branding role that "giga" played for Gigafactory, originally pitched as producing more batteries than the rest of the world combined.



Source: SemiAnalysis Accelerator Model, SemiAnalysis Foundry Model



Source: SemiAnalysis Wafer Fab Equipment Model, SemiAnalysis Foundry Model

Process IP is the deeper constraint and the reason the Samsung and TSMC deals matter most. Tesla has no manufacturing IP. The incumbents hold proprietary process technology. GAA transistor design, interconnect, lithography and etch recipes, and yield engineering refined over decades. Licensing is the only realistic path. Our view is that Terafab, if it ever reaches volume, operates as an integration fab on a licensed node the way Rapidus is attempting, not as a greenfield process developer.

The memory claim is the hardest to square. Musk expressed his opinions in the Tesla Q4 2025 call that “memory is an even bigger limiter than AI logic” without specifying what type of memory - HBM, LPDDR, or NAND. Each is a distinct process with IP concentrated in Samsung, SK Hynix, and Micron and thousands of patents. Long-term supply contracts or co-investment with an existing DRAM maker is the realistic path, not fabrication from scratch.

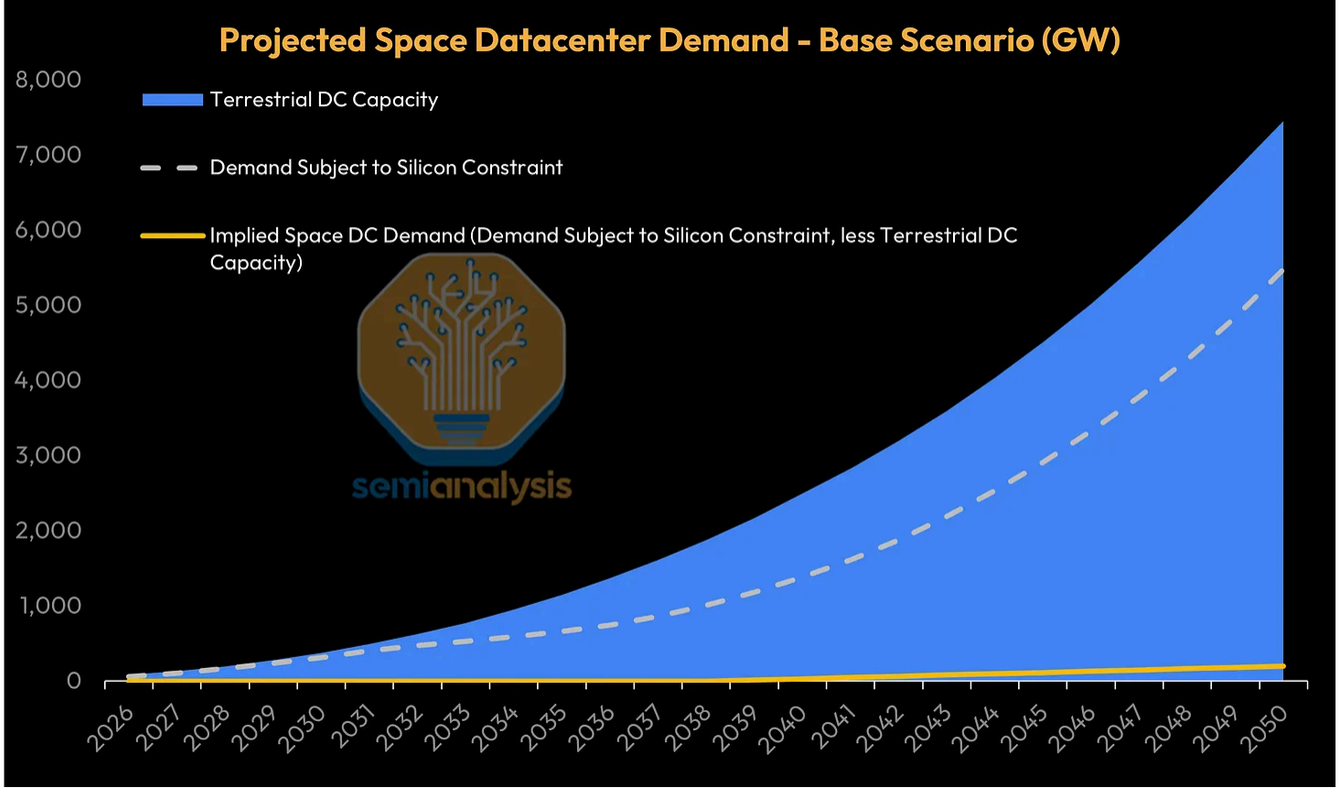
The SemiAnalysis AI Space Datacenter TCO Model allows users to bake Terafab capacity additions into the overall chip constraint model. Our Terafab base case doesn’t fully adopt the full extent of Elon Musk’s ambitious plans, but it still lifts the silicon constraint somewhat.

Source: SemiAnalysis AI Space Datacenter TCO Model

Assuming the aforementioned scenario adopted by our AI Space Datacenter TCO Model in its base case, whereby we depart from our other industry models to show accelerating incremental datacenter capacity additions and a meaningful step up in the pace of chip fab capacity addition, we see the following results.

With incremental datacenter capacity additions eventually in the hundreds of GW annually, and obstacles from gas turbine availability to EUV tool production constraints are removed, demand could be met by space or terrestrial datacenters.

Users of our AI Space Datacenter TCO Model can dynamically adjust these capacity parameters and silicon constraint assumptions to dial in scenarios as needed.



Source: SemiAnalysis AI Space Datacenter TCO Model

In such a scenario, the question then shifts from whether Earth or Space has more datacenter capacity, to a question of the total cost of ownership in Earth versus Space - bringing us to Part Three of this article.

Part Three: The Total Cost of Ownership Framework for Space Datacenters

In our total cost of ownership framework, we separate costs into three distinct buckets: **IT Capital Costs**, **Datacenter Capital Costs**, and **Operating Costs**. We begin with a line-by-line build of cluster capital requirements and then roll those upfront costs together with recurring operating expenses and an amortized cost of the datacenter to produce the below summary cost of ownership in $/hr/GPU and ultimately Levelized Cost of Compute expressed in units of $/hr per PFLOP.

Looking at a 30.5kW B300 datacenter deployed in 2026, we see that the total monthly cost of ownership is nearly four times higher for a space deployment as compared to a terrestrial deployment. Levelized Cost of Compute (LCOC) measured in $/PFLOP-hour for space is well over 4x the cost for an Earth datacenter. This difference is entirely driven by the much higher datacenter capex cost - 8x higher for space datacenters, but with a monthly cost of ownership 17x higher given the 5y useful life for space datacenters vs a 15Y useful life for Earth datacenters.

We use B300s and contemporary mainstream GPU as references for TCO analysis for the next few years, **but it is much more likely that smaller, efficient and specialized chips akin to Tesla's FSD chips will actually be deployed**.

AI Cloud Total Cost of Ownership Summary: Space vs Terrestrial (Extended Prices)	Unit	B300 1200W (Orbital)	B300 1200W
Site		Space	Terrestrial
Customer Profile		Hyperscaler	Hyperscaler
Year		2026	2026
Program Power	W	30,565	30,565
Power per GPU	W	1,910	1,910
GPUs	GPUs	16	16
GPU per Server	GPUs	8	8
IT Cluster Capital Cost	USD	$980,882	$986,158
Datacenter Capital Cost[1]	USD	$3,086,332	$382,061
Total Program Capital Cost	**USD**	**$4,067,215**	**$1,368,219**
Weighted Average Cost of Capital	%	15.0%	10.3%
Datacenter Useful Life in Years	Years	5	15
Monthly IT Cluster Capital Cost of Ownership	USD	$23,335	$21,099
Monthly Datacenter Capital Cost of Ownership[1]	USD	$73,424	$4,176
Monthly Operating Cost of Ownership	USD	$4,167	$2,449
Total Monthly Cost of Ownership	**USD**	**$100,925**	**$27,724**

1. Includes launch cost.

Source: SemiAnalysis AI Space Datacenter TCO Model

AI Cloud Total Cost of Ownership Summary: Space vs Terrestrial (per GPU-hr)	Unit	B300 1200W (Orbital)	B300 1200W
Site		Space	Terrestrial
Customer Profile		Hyperscaler	Hyperscaler
Year		2026	2026
IT Capital Cost of Ownership	USD/hr/GPU	$2.00	$1.81
Datacenter Capital Cost of Ownership	USD/hr/GPU	$6.29	$0.36
Operating Cost of Ownership	USD/hr/GPU	$0.36	$0.21
Total Cost of Ownership, Pre-SLA	**USD/hr/GPU**	**$8.64**	**$2.37**
Datacenter Capital Cost as % of TCO	*%*	*72.8%*	*15.1%*
Radiation Availability	%	95%	100%
Additional % GPUs needed for 99% SLA	%	20%	5%
IT Capital Cost of Ownership, Post-SLA	USD/hr/GPU	$2.52	$1.90
Datacenter Capital Cost of Ownership, Post-SLA	USD/hr/GPU	$7.94	$0.38
Operating Cost of Ownership, Post-SLA	USD/hr/GPU	$0.45	$0.22
Levelized Cost of Compute (LCOC)[1]	**USD/hr/GPU**	**$10.91**	**$2.49**
Marketed PFLOPS (FP4)	PFLOPS/GPU	15.0	15.0
Inference Throughput[2]	Tok/s/GPU	5,133.0	5,133.0
LCOC per PFLOP-hour	$/PFLOP-hr	$0.73	$0.17
LCOC per B Tokens	$/B tokens	$590.67	$134.87

1. Levelized Cost of Compute; refers to total cost of ownership after adjusting for radiation availability and additional GPUs needed.

2. DeepSeek R1 FP4. Uses 8k input, 1k output and 1k input, 1k output tokens 50/50 mix, 100 interactivity.

Source: SemiAnalysis AI Space Datacenter TCO Model

Let's step into each individual category and examine key cost drivers.

For **IT Capital Cost of Ownership**, space and terrestrial datacenters are largely identical. Servers, networking fabric, and the software, storage and orchestration stack are the same in orbit as on the ground with the exception of warranty and capitalized burn-in costs.

Turning to **Datacenter Capital Cost of Ownership**, the space and terrestrial models diverge most sharply. Terrestrial datacenters' largest cost buckets are shell construction, chillers and cooling towers, transformers and grid interconnects, and physical security. Orbital deployments eliminate land entirely and replace each of these line items with space-specific equivalents which we list below.

When analyzing **Operating Cost of Ownership**, the cost profile inverts. Terrestrial datacenters pay for grid power and run a steady stream of on-site technicians for maintenance, while space datacenters do not have an incremental cost of power (they get power from solar panels) and they pay ground control operating costs which are allocated to the satellite.

AI Cloud Total Cost of Ownership Framework

Line Item	Space	Terrestrial
	IT Capital Cost of Ownership	
Servers	Standard GPU servers (e.g., B300s)	Same
Networking	NVLink / InfiniBand / Ethernet	Same
Software, storage, others	Storage, orchestration, installation	Same
Server service	-	Recognized upfront
Capitalized burn-in	Recognized upfront	-
	Datacenter Capital Cost of Ownership	
Cooling / HVAC	-	Standard requirement for terrestrial
Core / shell	-	Standard requirement for terrestrial
Pod / fit-out	-	Standard requirement for terrestrial
Land	-	Significant cost in Tier 1 markets
Shielding	Protects electronics from radiation and debris	-
Solar array	Generates electrical power in orbit	-
Radiator hardware	Rejects heat, improves with breakthroughs	-
Cold plates	Transfers heat from GPUs	-
Pump and heat exchanger	Circulates coolant	-
Battery	Powers satellite during eclipse	-
Comms hardware	Onboard antennas and laser links	-
Ground segment	Earth-based stations and operations	-
Bus hardware	Computers, sensors, power distribution	-
Propulsion hardware	Thrusters for maneuvering and deorbit	-
Feed system	Plumbing delivering propellant to thrusters	-
Propellant + tank	Fuel and storage container	-
Structure and integration	Physical frame	-
Assembly, integration, testing	Building and verifying the satellite	-
Launch cost	Significant cost, declines over time	-
	Operating Cost of Ownership	
Electricity cost	-	Grid electricity, onsite gas
Remote hands + support	-	On-site technicians, maintenance
Internet connection	-	Standard requirement for terrestrial
Operations	Scales with mission value	Standard requirement for terrestrial

Source: SemiAnalysis AI Space Datacenter TCO Model

IT Capital Cost of Ownership

For the first few years of space datacenter deployment, we expect that IT equipment deployed will be fundamentally similar, with a low degree of customization and similar costs. Over time, if the pace of space datacenter deployments increase, we could expect to see some limited divergence in cost, though with AI XPUs, much of the IT cost is related to the core GPU+HBM package, leaving limited room for variance due to customization.

Using our 2026 30.5kW Datacenter concept, capital costs for a B300 cluster include the server cost itself, scale-out networking, attached storage and several smaller line-item costs required to bring a cluster online. Two servers consisting of 8 B300s each carries a base server cost of $880,400 with Service, Networking, Storage, Software and other infrastructure components contributing an additional $89,956. With an additional capitalized burn-in cost of $10,526 for space deployments, and a capitalized server service cost of $15,802 for terrestrial deployments, this brings the total upfront IT cluster capex to $980,882 for space deployments and $986,158 for terrestrial deployments.

To reflect the increased risk from the relative immaturity of space datacenter deployments, we use a flat Weighted Average Cost of Capital (WACC) of 10.3% for terrestrial deployments, which corresponds to a ~7% pre-tax cost of debt, a 20% cost of equity, and a 75/25 debt-equity split. This compares to an initial WACC of 15.0% for space deployments that declines over time to reach parity at 10.3% within ~10 years once space datacenters are more mature and de-risked. Assuming a 5-year useful life for IT equipment, this translates into an IT capital cost of ownership of approximately $2.00/hr/GPU for space versus $1.81/hr/GPU for terrestrial.

IT Capital Cost of Ownership Summary

	Unit	B300 1200W (Orbital)	B300 1200W
Site		Space	Terrestrial
Customer Profile		Hyperscaler	Hyperscaler
Year		2026	2026
Cluster Capital Costs			
GPU Server Cost	USD	$880,400	$880,400
Server Service	USD	$0	$15,802
Networking Cost	USD	$65,856	$65,856
Storage Cost	USD	$9,770	$9,770
Software License and Other Costs	USD	$6,230	$6,230
Other Installation	USD	$8,100	$8,100
Capitalized Burn-in Cost	USD	$10,526	$0
Total Upfront IT Cluster Capex	**USD**	**$980,882**	**$986,158**
Weighted Average Cost of Capital	%	15.0%	10.3%
Useful Life in Years	Years	5	5
Monthly IT Cluster Capital Cost of Ownership	**USD/mth**	**$23,335**	**$21,099**
IT Capital Cost of Ownership	**USD/hr/GPU**	**$2.00**	**$1.81**

Source: SemiAnalysis AI Space Datacenter TCO Model

Minor adaptations include tolerating radiation-induced faults such as Single Event Upsets (SEUs) or Single Event Functional Interrupts (SEFIs) through Error Correcting Code (ECC) memory, watchdog resets, and graceful restart mechanisms, an approach already demonstrated at scale by constellations such as Starlink which did not require expensive rad-hard processors.

In practice, the compute nodes themselves remain fundamentally the same as their terrestrial counterparts, but the surrounding power, thermal, and reliability systems are engineered so that standard datacenter hardware can operate reliably in the radiation, thermal, and power environment of orbit. We cover these costs in the following sections.

Datacenter Capital Cost of Ownership

While IT Capital Costs for space datacenters and Earth datacenters are similar, the cost gap is much wider when we look at Datacenter Capital Cost of Ownership. For terrestrial datacenters, there are two ways to incorporate datacenter facility costs, an opex-based approach and a capex-based approach.

In the opex based cost model, the Hyperscaler or Neocloud rents colocation capacity and cost is quoted in terms of USD per kilowatt-hour (kWh) of critical IT power per month. In the capex-based model, we assume the Hyperscaler or Neocloud builds and

owns the facility outright, and we calculate a levelized cost of ownership based on the following: 5-year useful space datacenter facility life until 2032, then assuming improvements drive this to 10-year useful life after.

The AI Space Datacenter TCO Model exclusively uses the capex-based model for Earth datacenters to allow a like for like comparison as we think the most likely business model will be a vertically integrated deployment where a provider like SpaceX will own the space datacenter as well as the IT compute installed within. Further breakdowns of the useful life and capex required for self-built Earth datacenters can be found in our earlier articles on datacenter energy and cooling.

Compared to Earth datacenters, orbital deployments eliminate land entirely and replace each of the usual Earth datacenter capex line items with space-specific equivalents: satellite structure and shielding instead of buildings, radiators, cold plates and closed-loop liquid cooling systems instead of HVAC, solar panels and batteries instead of grid electrical, and the satellite bus (ADCS, propulsion, command and data handling) instead of facilities. On top of that, space carries key cost lines that simply do not exist terrestrially — launch ($/kg to orbit), one of the major cost driver of the entire model, as well as other key costs covering radiation shielding, propulsion and assembly, integration and test.

For our 30.5kW B300 datacenter deployed in 2026, overall datacenter capital costs (this excludes IT costs) at today's technology come out at $3.1M for a space deployment versus $382K for a terrestrial deployment. Space datacenters deployed in 2026 are more costly because of a larger upfront capital cost of deployment, with the largest driver being launch costs at $1.6M out of the total ~$3.1M datacenter capital cost.

The cost difference is even starker when considering levelized datacenter costs - because space datacenters are expected to have a useful life of only 5 years (this is assumed to eventually extend to 10 years from 2032 onwards due to improvements in in-space robotics) vs the standard 15 years for Earth-based datacenters (due to buildings and facilities that outlast the GPUs themselves). When these are taken into account, datacenter capital costs of ownership are a whopping 17x higher than for terrestrial datacenters at $6.29/hr/GPU for space compared to $0.36/hr/GPU!

Within our AI Space Datacenter TCO Model, most of these line items are fully adjustable, allowing users to customize individual items based on their assumptions around cost and technology scaling for these space systems.

Datacenter Capital Cost of Ownership Summary			
	Unit	B300 1200W (Orbital)	B300 1200W
Site		Space	Terrestrial
Customer Profile		Hyperscaler	Hyperscaler
Year		2026	2026
Terrestrial Datacenter Costs			
Power, Transformers, Distribution	USD		$170,378
Cooling, HVAC	USD		$100,678
Core, Shell Facilities	USD		$51,630
Pod, Fit Out	USD		$20,652
Land	USD		$38,722
Terrestrial Datacenter Capital Costs	**USD**		**$382,061**
Space Datacenter Costs			
Shielding	USD	$41,358	
Solar array	USD	$198,574	
Radiator hardware	USD	$136,205	
Cold plates	USD	$12,480	
Pump and heat exchanger	USD	$30,565	
Battery	USD	$874	
Comms. hardware	USD	$95,000	
Communications Ground segment (allocated)	USD	$24,000	
Bus hardware	USD	$176,841	
Propulsion hardware	USD	$400,000	
Feed system	USD	$32,000	
Propellant + tank	USD	$25,325	
Structure and integration	USD	$105,244	
Assembly, integration, and testing (AIT)	USD	$188,170	
Launch cost	USD	$1,619,696	
Space Datacenter Capital Costs	**USD**	**$3,086,332**	
Weighted Average Cost of Capital	%	15.0%	10.3%
Useful Life in Years	Years	5	15
Monthly Datacenter Capital Cost of Ownership	**USD/mth**	**$73,424**	**$4,176**
Datacenter Capital Cost of Ownership	**USD/hr/GPU**	**$6.29**	**$0.36**

Source: SemiAnalysis AI Space Datacenter TCO Model

Operating Cost of Ownership

As mentioned above, when it comes to operating cost of ownership, the TCO profile inverts between space and terrestrial. Terrestrial datacenters run a steady stream of on-site technicians, grid power tariffs (electricity and onsite gas), and continuous hardware replacement and cooling upkeep.

The orbital model has none of these recurring costs: there is no servicing once in orbit, no ongoing power bill because solar capex is paid upfront, and failures are absorbed through roughly 20% spare redundancy rather than physical maintenance. The orbital model does have several opex lines related to ground operations cost (launch, control, communications) that do not apply to a terrestrial model.

Turning back to our Earth-based B300 cluster, we assume an electricity cost of $0.087/kWh, a power utilization rate of 80% typically achieved by serving steady training workloads and a Power Usage Effectiveness (PUE) of 1.35. PUE represents the additional power above Critical IT Power required for cooling, power delivery, and other facility-level systems. Total operating cost of ownership is dominated by power tariff and lands at $0.21/hr/GPU.

Operating costs for space datacenters are dominated by allocated operations costs, at $4,167 per month, amounting to $0.36/hr/GPU. This scales rapidly however, as fixed ground operations costs are amortized over a larger fleet of space datacenters - our

modeling shows operating costs for space datacenters dropping as low $0.15/hr/GPU by the mid-2030s - one third to half of the operating cost of Earth datacenters, though it's important to remember that IT and datacenter capital cost is still the dominant cost component for both Earth and Space.

AI Cloud Operating Cost of Ownership Summary			
	Unit	B300 1200W (Orbital)	B300 1200W
Site		Space	Terrestrial
Customer Profile		Hyperscaler	Hyperscaler
Year		2026	2026
Operating Costs			
Electricity Cost	USD/kWh/mth	$0.00	$0.09
Utilization Rate	%	95%	80%
Power Usage Effectiveness (PUE)	Ratio	1.05	1.35
All-in Power Consumption	kW	30.56	30.56
Total Power Costs per Month	**USD/mth**	**$0**	**$2,068**
Remote Hands + Support Engineer	USD/mth	$0	$262
Internet Connection	USD/mth	$0	$78
Operations	USD/mth	$4,167	$41
Monthly Operating Cost of Ownership	**USD/mth**	**$4,167**	**$2,449**
Operating Cost of Ownership	**USD/hr/GPU**	**$0.36**	**$0.21**

Source: SemiAnalysis AI Space Datacenter TCO Model

From Total Cost of Ownership (TCO) to Levelized Cost of Compute (LCOC)

Taking our per-GPU-hour costs for IT capital costs, datacenter capital costs, and operating costs together, we see that total cost of ownership per GPU-hour a B300 datacenter is $8.64/hr/GPU for an orbital deployment compared to $2.37/hr/GPU for a terrestrial deployment.

The final adjustments required to bridge the total cost of ownership to a levelized cost of compute (LCOC) stem from radiation availability and additional GPUs required. Radiation availability refers to compute availability net of what is temporarily degraded or affected by solar radiation, i.e. occasional faults that take the affected hardware offline temporarily - we model this at 95% for space versus 100% (no effect) for terrestrial. Additional GPUs required to reach a 99% SLA refers to additional GPUs provisioned for redundancy given hardware failures that cannot be fixed through software or a soft reboot - on earth we see cold spares of up to 5%, however given that orbital chips are not able to be mechanically repaired, we assume a 20% redundancy for failures.

This gives us a Levelized Cost of Compute (LCOC) in 2026 for our B300 cluster of $10.91/hr/GPU for an orbital deployment compared to $2.49/hr/GPU for a terrestrial deployment.

AI Cloud Total Cost of Ownership Summary: Space vs Terrestrial (per GPU-hr)

	Unit	B300 1200W (Orbital)	B300 1200W
Site		Space	Terrestrial
Customer Profile		Hyperscaler	Hyperscaler
Year		2026	2026
IT Capital Cost of Ownership	USD/hr/GPU	$2.00	$1.81
Datacenter Capital Cost of Ownership	USD/hr/GPU	$6.29	$0.36
Operating Cost of Ownership	USD/hr/GPU	$0.36	$0.21
Total Cost of Ownership, Pre-SLA	**USD/hr/GPU**	**$8.64**	**$2.37**
Datacenter Capital Cost as % of TCO	*%*	*72.8%*	*15.1%*
Radiation Availability	%	95%	100%
Additional % GPUs needed for 99% SLA	%	20%	5%
IT Capital Cost of Ownership, Post-SLA	USD/hr/GPU	$2.52	$1.90
Datacenter Capital Cost of Ownership, Post-SLA	USD/hr/GPU	$7.94	$0.38
Operating Cost of Ownership, Post-SLA	USD/hr/GPU	$0.45	$0.22
Levelized Cost of Compute (LCOC)[1]	**USD/hr/GPU**	**$10.91**	**$2.49**
Marketed PFLOPS (FP4)	PFLOPS/GPU	15.0	15.0
Inference Throughput[2]	Tok/s/GPU	5,133.0	5,133.0
LCOC per PFLOP-hour	$/PFLOP-hr	$0.73	$0.17
LCOC per B Tokens	$/B tokens	$590.67	$134.87

1. Levelized Cost of Compute; refers to total cost of ownership after adjusting for radiation availability and additional GPUs needed.

2. DeepSeek R1 FP4. Uses 8k input, 1k output and 1k input, 1k output tokens 50/50 mix, 100 interactivity.

Source: SemiAnalysis AI Space Datacenter TCO Model

However - the silicon area per GPU and thus compute throughput and IT power per GPU continues to evolve. The trend is for a greater number of individual silicon dies per GPU with more compute dies and input-output (I/O) dies finding their way into each contemporary GPU. However - the trend may be the opposite for AI Chips deployed into space datacenters. **It is much more likely that smaller, efficient and specialized chips akin to Tesla's FSD chips will actually be deployed**. Larger chips will have a greater compute throughput in PFLOPs and produce more tokens/s, but ultimately both smaller efficient space AI chips and larger ground GPUs will have capabilities that scale by power.

To normalize for this and model in a way that is agnostic to the definition of a GPU or the size of an AI chip, we project PFLOPs/Watt, Capex/Watt, and Watt per silicon area. As a result, our figure of merit for measuring LCOC in the long term is LCOC per PFLOP-hour - the cost per PFLOP of compute throughput for one hour.

On a marketed FP8 dense FLOPS basis, B300s are rated at 4,500 dense FP4 TFLOPS, giving ground-based deployments an LCOC of $0.17/PFLOP-hr per marketed FP4 Dense PFLOP, and space datacenters an LCOC of $0.73/PFLOP-hr.

Similarly, our open benchmarks under InferenceX suggest that with disagg TRT, MTP using Deepseek R1, B300 token throughput under FP4 is ~5,100 Tok/s per GPU, mapping to an inference LCOC of $590 per Billion Tokens for a 2026 space datacenter vs only $135 per Billion tokens for an Earth datacenter.

AI Cloud Performance per TCO			
	Unit	B300 1200W (Orbital)	B300 1200W
Site		Space	Terrestrial
Customer Profile		Hyperscaler	Hyperscaler
Year		2026	2026
Levelized Cost of Compute (LCOC)[1]	USD/hr/GPU	$10.91	$2.49
Logical GPUs per Server	Logical GPUs	8	8
Maximum Scale-up World Size	Logical GPUs	8	8
Marketed FLOPS across World Size (FP4)	PFLOPS	120.0	120.0
Effective Training FLOPS across World Size (FP4)	PFLOPS	60.0	60.0
Aggregate Memory Bandwidth across World Size	TB/s	64.0	64.0
Marketed TFLOPS (FP4)	TFLOPS/GPU	15,000.0	15,000.0
Effective TFLOPS (FP4)	TFLOPS/GPU	7,500.0	7,500.0
Inference Throughput[2]	Tok/s/GPU	5,133.0	5,133.0
Memory Bandwidth per Logical GPU	TB/s/GPU	8.0	8.0
Marketed TFLOPS (FP4) / Memory Bandwidth	TFLOPS/TB/s/GPU	562.5	562.5
LCOC per PFLOP	$/hr per PFLOP	$0.73	$0.17
LCOC per Effective PFLOP	$/hr per PFLOP	$1.46	$0.33
LCOC per B Tokens	$/B tokens	$590.67	$134.87
LCOC per Memory Bandwidth	$/hr per TB/s	$1.36	$0.31

1. Levelized Cost of Compute; refers to total cost of ownership after adjusting for radiation availability and additional GPUs needed.

2. DeepSeek R1 FP4. Uses 8k input, 1k output and 1k input, 1k output tokens 50/50 mix, 100 interactivity.

Source: SemiAnalysis AI Space Datacenter TCO Model

As we have alluded to many times above, deploying a few 30kW space datacenters in 2026 is clearly sub-scale. For the space-ground cost parity and eventual crossover to happen within a commercially meaningful planning horizon, space LCOC needs to fall much faster than that of terrestrial datacenters.

So When Does Space Actually Get Interesting?

Space compute becomes interesting when it becomes economically viable by being at or near cost parity with terrestrial datacenters. There is still a world of engineering and cost challenges between now and then that SpaceX, the supply chain, and the space compute ecosystem will have to go through iterations of trial, error, and operationalization before cost parity is achieved.

Despite these challenges, the economic opportunity to meet exponential AI demand at a more cost-effective unit basis over the long-term is the north star for SpaceX, and something that operators, investors, and enthusiasts must not ignore.

In the next sections, we'll work through the base case and 'Elon Musk' case and how cost parity evolves differently between the two cases. We will dive into an explanation of each major space datacenter cost item, deriving these costs from first principles and how they evolve over the years. We will also discuss key considerations regarding launch costs, system and subsystem capital and operating costs, compute payload costs and more. Subscribers to our AI Space Datacenter TCO Model also have full access to the above calculations and assumptions and can dynamically adjust almost all of the underlying assumptions.

Space Compute Prerequisites and Obstacles

Deploying datacenters in space entails a very different, and as of today, much higher total cost structure. The other major obstacle to overcome stems from chip reliability and servicing.

Accounting for failure: At steady state operations, 3-6% is the estimated "mortality rate" of GPUs in ground datacenters today - this rate represents failures that require physical human intervention to resolve. This is practically impossible to replicate in space under today's economics, though robotics could be one future solution. To solve for this, our model assumes an over-provisioning of GPUs by 20%, and we also assume a capitalized "burn-in" cost to reflect the burn-in cycles done prior to launch to screen early mortality rates (estimated to 3-4x higher at 10-20% during burn-in)

Servicing and replacement challenges: It is more cost-effective to have small failure domains in order to minimize blast radius. This may also preclude the use of larger world-size rack-scale servers like the GB300 NVL72 and the VR NVL72. Localized failure design coupled with over provisioning GPUs would enable the satellite to achieve near or at 100% operating capacity despite failures with no viable remediation.

Faults can be worked around and planned for: Faults include SEFI/latchup/SEU, account for resets as a part of operations. Operators will need to implement software redundancies to mitigate the radiation effects caused by exposure.

There are further arguments challenging the notion of space compute, though as with all space compute modeling, no one knows for certain and visibility into these topics is generally limited.

Operating datacenters in orbit stacks the failure modes, (i.e., radiation failure layered on top of the same failure modes that ground GPUs face). A bricked firmware update, thermal cycling, memory errors, radiator deployment issues, or (as mentioned) radiation induced failures can cause assets to fail without the convenience of having a ground datacenter crew to swap the chip within the hour. SpaceX has the only meaningful operational dataset on commercial silicon in LEO, limiting true visibility on what the operational issues and learnings would be.

There are also the Kessler constraints. SSO orbital slots are very scarce, and SpaceX isn't the only operator sending satellites into orbit. The Chinese space race is accelerating, and it is unclear what collaboration is going to look like over orbital real estate. This presents a major risk in terms of collisions and orbital debris, which is difficult to ascertain given the early stages of the next satellite ramp era.

Though many future technological obstacles remain, there are present-day mitigation techniques that could still provide a path to early generation platforms to drive operational learnings from.

Consider a hypothetical space datacenter with many independent B300 NVL8 islands (with 8 GPUs each). NVLink can be used within the island as per normal (fast

collective ops and parallelism). A scale-out network can then be used between these islands, handling request routing, replication, checkpointing and MoE Routing. A front-end fabric can also be deployed for data ingress and egress, storage access and fleet management. Like its terrestrial counterpart, link distances are short and fixed, power is routed such that a failure doesn't drop the whole domain, and thermal is sized so the rack can run at a stable temperature across orbital seasons/eclipses. Deploying a rack-scale server with a larger world size of 72 GPUs for instance would require recreating a rack-like architecture in a way that is compatible with the constraints for space datacenters.

As the fleet of deployed space datacenters expands over time, as cost-downs and efficiency gains of critical components are achieved, and as launch costs decline (via Starship platform), scale benefits begin to emerge, driving Levelized Cost of Compute for Space vs Earth closer to parity.

Scenario Analysis: The Base Case and the 'Elon Musk' Case

Terrestrial datacenter costs are rising. Electricity has climbed 2–5%/year in major datacenter markets, with Virginia spiking 15–20% in 2023–2024 from capacity strain. Facility construction is projected at $15–18M/MW by 2027 as demand outstrips build capacity. Land and permitting timelines have stretched from 12–18 months to 36–48 months in constrained markets.

Meanwhile, space inputs are falling. Starship targets $250–500/kg by 2028, and long-run costs below $185/kg, though our base case uses more conservative (i.e. higher) launch cost assumptions. Starlink has demonstrated more than an order of magnitude in bus cost reduction over heritage satellites through vertical integration and volume. Silicon solar arrays at Starlink volume hit 30-100X cost reduction for legacy GaAs.

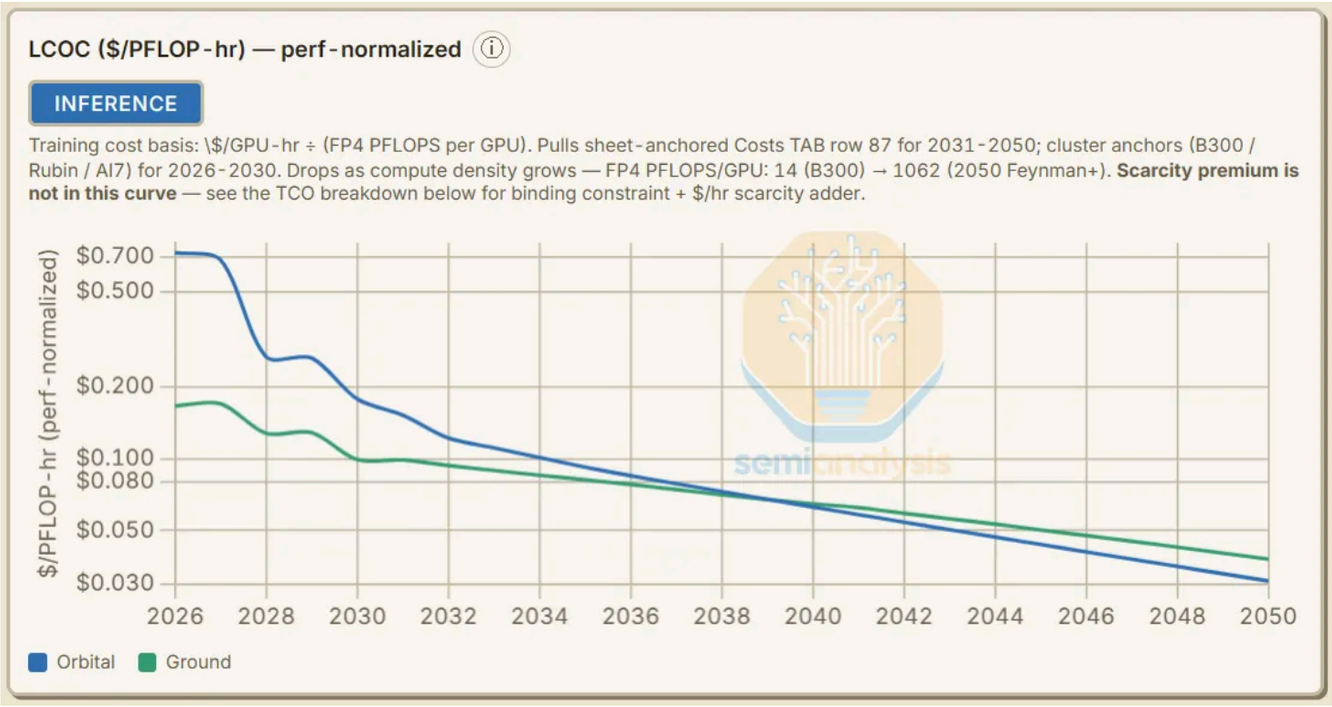
It is clear that space-ground cost parity is a medium- to long-term future goal, perhaps a decade or more from now. Uncertainty remains over how these trends will develop and how engineers overcome technological obstacles.

We present two scenarios here - (1) a base case (albeit reasonably optimistic), and (2) the Elon Musk Scenario.

In the base case, we assume that critical roadblocks like radiation availability and chip reliability are largely solved (or at least provisioned for), and that key cost scaling breakthroughs for radiators, launch costs and solar costs are achieved. The base case also takes a bullish view on chip production ramps, Earth datacenter capacity addition (though this is accomplished via significant cost inflation) as well as overall AI demand.

Our datacenter estimates are higher than that of our Datacenter Model because while the datacenter model focuses mostly on confirmed datacenters and confirmed land

bank, we take a more aspirational and optimistic view on datacenter pipelines. **In this scenario - space-earth total cost of ownership parity is reached by ~2040, and even comes close by the early 2030s - with space only 30% more expensive than Earth-based datacenters.** This might even open the door to deployments as soon as the turn of the decade if capacity tightness is severe enough on earth.



Source: SemiAnalysis AI Space Datacenter TCO Model

The Elon Musk scenario assumes that adding incremental terrestrial datacenter capacity is much more difficult, and overall supply scarcity pushes Earth datacenter capex costs far higher - to nearly $55M/MW vs the $12-14M/MW modeled today. We also assume advances in cost scaling push space datacenter capex per MW down to $11M/MW. As an important note, this is considered as a large upside case and not our official, unified view on terrestrial supply over the long-run. Silicon production capacity is helped by Terafab's additional 1,000kWSPM by 2040, though both the Base Case and the Elon Musk Scenario envision a massive expansion of chip fab capacity over the long-run.

There are two important outcomes from the Elon Musk scenario. First, scarcity of terrestrial datacenter supply pushes costs up to the extent that the gap between Space and Earth datacenter capex per MW more than triples and more importantly, there is far less terrestrial datacenter capacity added, leading to an order-of-magnitude greater spillover demand that can only be fulfilled by space datacenters.

Parity Scenario	Units	2026	2039	
		Current Costs	Base Case	Elon Musk Scenario
Unconstrained AI Critical IT Power Demand	GW	60.0	3,369.5	3,369.5
Silicon Production Limit	GW	60.0	1,184.4	1,184.4
Constrained AI Critical IT Power Demand	GW	60.0	1,184.4	1,184.4
Terrestrial Datacenter Global Critical IT Power	GW	89.0	2,172.2	741.6
Space Datacenter Critical IT Power Demand	**GW**	**0.0**	**0.0**	**442.8**
IT Capex per MW (similar for Space and Earth)	$M/MW	$32.1	$46.2	$46.2
Earth-based Datacenter Capex per MW	$M/MW	$12.5	$34.6	$53.4
Space Datacenter Capex per MW	$M/MW	$101.0	$15.8	$11.0

Source: SemiAnalysis AI Space Datacenter TCO Model

In the Elon Musk scenario, meeting this spillover demand using space datacenters is the more cost-effective solution. Space DC costs are nearly 20% lower than Earth DC costs by 2039, while in our base case space and earth only just reach parity then.

Space Datacenter Cost Scenario Analysis	Units	2026 Current Costs		2039 Base Case		2039 Elon Musk Scenario	
		Space	Terrestrial	Space	Terrestrial	Space	Terrestrial
IT Cluster Capital Cost of Ownership	$/hr/GPU	$2.00	$1.81	$11.40	$11.46	$11.40	$11.46
Datacenter Cluster Capital Cost of Ownership	$/hr/GPU	$6.29	$0.36	$2.44	$4.37	$1.69	$6.73
Operating Cost of Ownership	$/hr/GPU	$0.36	$0.21	$0.27	$1.11	$0.27	$1.11
Total Cost of Ownership	**$/hr/GPU**	**$8.64**	**$2.37**	**$14.11**	**$16.94**	**$13.36**	**$19.30**
Radiation Availability	%	95.0%	100.0%	95.0%	100.0%	95.0%	100.0%
Additional GPUs for 99% SLA	%	20.0%	5.0%	20.0%	5.0%	20.0%	5.0%
Levelized Cost of Ownership (LCOC)	$/PFLOP-hr	$0.73	$0.17	$0.068	$0.067	$0.064	$0.077
Parity Ratio - Space LCOC / Earth LCOC	x	4.38x		1.00x		0.83x	

Source: SemiAnalysis AI Space Datacenter TCO Model

The Elon Musk scenario assumes transformational improvement in space launch costs - only $80/kg vs the ~$1,700/kg today and the $485/kg in our base case. This scenario sees a drop in total space datacenter program cost per watt from $133/W today to $57/W. However, it is interesting to observe that despite the Elon Musk case assuming launch costs 85% below that of our base case, total program upfront capital cost is only 8% lower and LCOC only drops by 6%.

For anyone who has studied total cost of ownership for contemporary terrestrial datacenters, the reason for this is clear: IT capital cost of ownership for today's datacenters is 75-80% of total cost of ownership. This is due to a high absolute capital cost as well as a lower useful life - we assume 5 years for space datacenters before 2032 - 10 years for space datacenters deployed in 2032 and after vs 15 years for terrestrial datacenters. A 20% reduction in space datacenter capital cost only results in a 5% drop in total cost of ownership.

An important note on modeling cost of capital for our analysis: we use a flat Weighted Average Cost of Capital (WACC) of 10.3% for terrestrial deployments, which corresponds to a ~7% pre-tax cost of debt, a 20% cost of equity, and a 75/25 debt-equity split. This compares to an initial WACC of 15.0% for space deployments that declines over time to reach parity at 10.3% within ~10 years.

We acknowledge this sits above SpaceX's current actual cost of capital. Per the S-1, the $20B SpaceX Bridge Loan carries an effective rate of 4.58% (SOFR + 75-175 bps), and

equipment leases via Valor run at ~5.5% blended, implying a weighted average pre-tax cost of debt of ~4.86%. However, we deliberately apply a higher cost of capital at the outset because:

- (a) orbital AI compute is a new, unproven industry with no comparable cash-flow track record to support debt at SpaceX corporate rates,
- (b) the SpaceX bridge loan was entered into in March 2026 (after the xAI merger in February 2026, which had acquired X even earlier in March 2025), stabilizing the legacy X and xAI term loans with enterprise-level cash flows,
- (c) project-level financing for first-of-its-kind infrastructure typically prices at a premium to corporate debt, and
- (d) the bridge loan matures in 2027-28, after which permanent financing may price higher in a normalized rate environment.

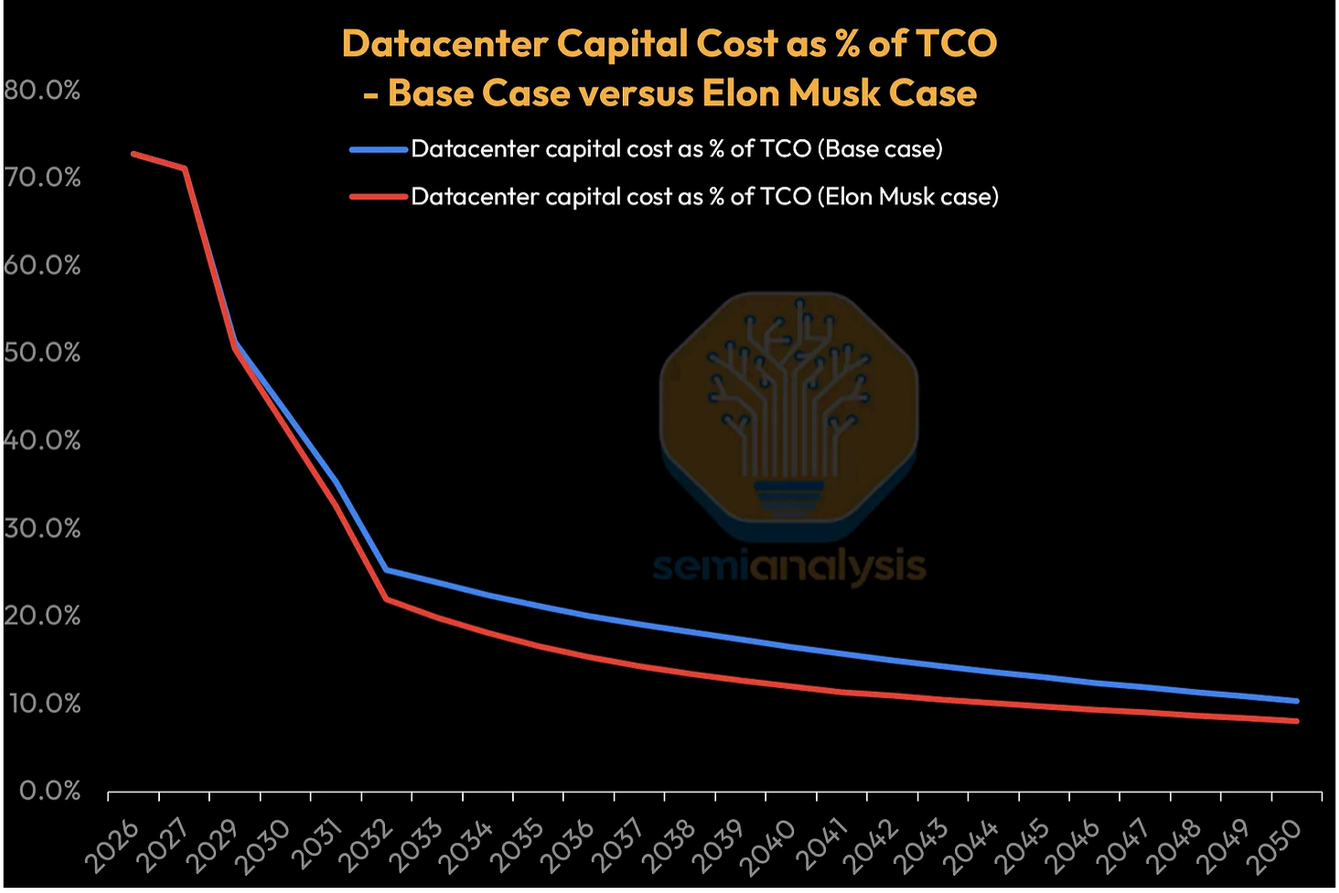
We believe this is more representative of standalone launch risks. In comparison, Hyperscaler WACC is typically 8-10% and pure project finance for established infrastructure assets is 5-8%.

Space Datacenter Cost Scenario Analysis

	Units	2026	2039	
		Current Costs	Base Case	Elon Musk Scenario
IT Cluster - % of Space DC + IT Cost	%	24.1%	74.5%	80.8%
Solar Array - % of Space DC + IT Cost	%	4.9%	4.2%	4.6%
Radiator - % of Space DC + IT Cost	%	3.3%	2.9%	3.1%
Bus Hardware - % of Space DC + IT Cost	%	4.3%	2.3%	2.5%
Propulsion Hardware - % of Space DC + IT Cost	%	9.8%	0.3%	0.4%
Launch costs - % of Space DC + IT Cost	%	39.8%	9.4%	1.7%
Launch Cost	$/kg	$1,748	$485	$81
IT Cluster Cost per Critical IT Power	$/W	$32.1	$46.2	$46.2
Solar Array Cost per Critical IT Power	$/W	$6.5	$2.6	$2.6
Radiator Cost per Critical IT Power	$/W	$4.5	$1.8	$1.8
Bus Hardware Cost per Critical IT Power	$/W	$5.8	$1.4	$1.4
Propulsion Hardware - Cost per Critical IT Power	$/W	$13.1	$0.2	$0.2
Launch costs - Cost per Critical IT Power	$/W	$53.0	$5.8	$1.0
All other Costs - Cost per Critical IT Power	$/W	$18.2	$4.0	$4.0
Total Cost per Critical IT Power	**$/W**	**$133.1**	**$62.0**	**$57.1**

Source: SemiAnalysis AI Space Datacenter TCO Model

A counterintuitive result of our analysis is that in 2039, launch costs don't move the needle that much. If space datacenters can scale to the point where they are nearly competitive with Earth datacenters, then collectively space datacenter costs will only be 15-20% of total cost of ownership. A key driver of scaling space datacenter costs in the first place is a huge drop in space launch costs - we see this dropping 10x in terms of $/W, a much faster decline than other cost line items - but this means that launch costs drop from 40% of space DC total program capex to only ~10% of space DC total program capex. A further 50% drop in a 10% line item does little to drive further declines in total cost of ownership.



Source: SemiAnalysis AI Space Datacenter TCO Model

This is why, as per the table below, the space-earth gap in 2039 does not move as meaningfully as one would imagine with respect to launch costs. Many other line items like radiators, solar arrays, bus hardware, are assumed to have also scaled in the meantime.

Space/Terrestrial LCOC Analysis for 2039 (Base Case)

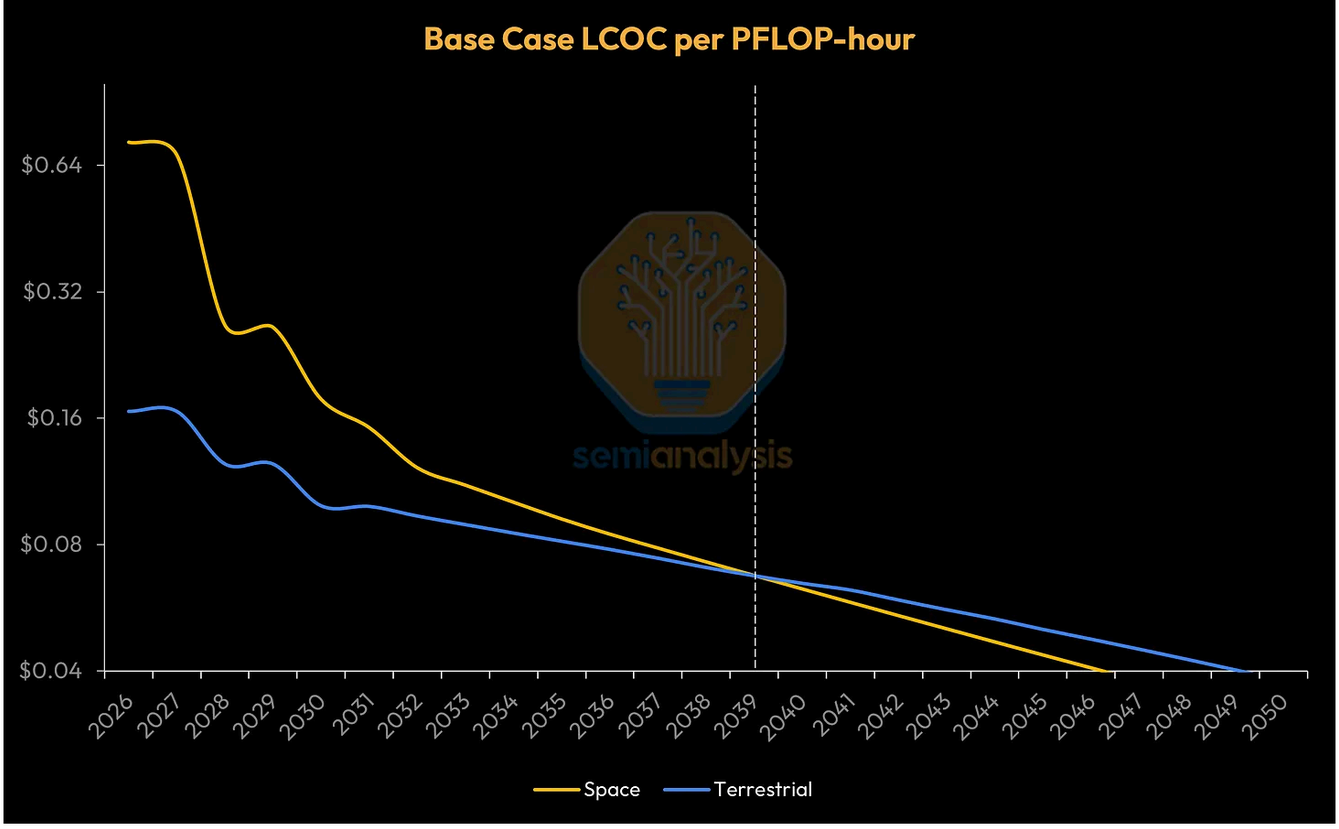
Terrestrial datacenter costs ($M/MW) \ Launch cost ($/kg)	50	100	200	400	485	600	800	1000	1200
25	1.02x	1.03x	1.04x	1.07x	1.08x	1.10x	1.12x	1.15x	1.18x
30	0.98x	0.99x	1.00x	1.03x	1.04x	1.05x	1.08x	1.11x	1.14x
35	0.94x	0.95x	0.96x	0.99x	1.00x	1.02x	1.04x	1.07x	1.10x
40	0.91x	0.91x	0.93x	0.95x	0.96x	0.98x	1.00x	1.03x	1.05x
45	0.88x	0.88x	0.90x	0.92x	0.93x	0.94x	0.97x	0.99x	1.02x
50	0.85x	0.85x	0.87x	0.89x	0.90x	0.91x	0.94x	0.96x	0.98x
55	0.82x	0.83x	0.84x	0.86x	0.87x	0.88x	0.91x	0.93x	0.95x
60	0.79x	0.80x	0.81x	0.83x	0.84x	0.86x	0.88x	0.90x	0.92x

Green font indicates scenarios where space LCOC per PFLOP-hr is cheaper than terrestrial.
Red font indicates scenarios where space LCOC per PFLOP-hr is more expensive than terrestrial.

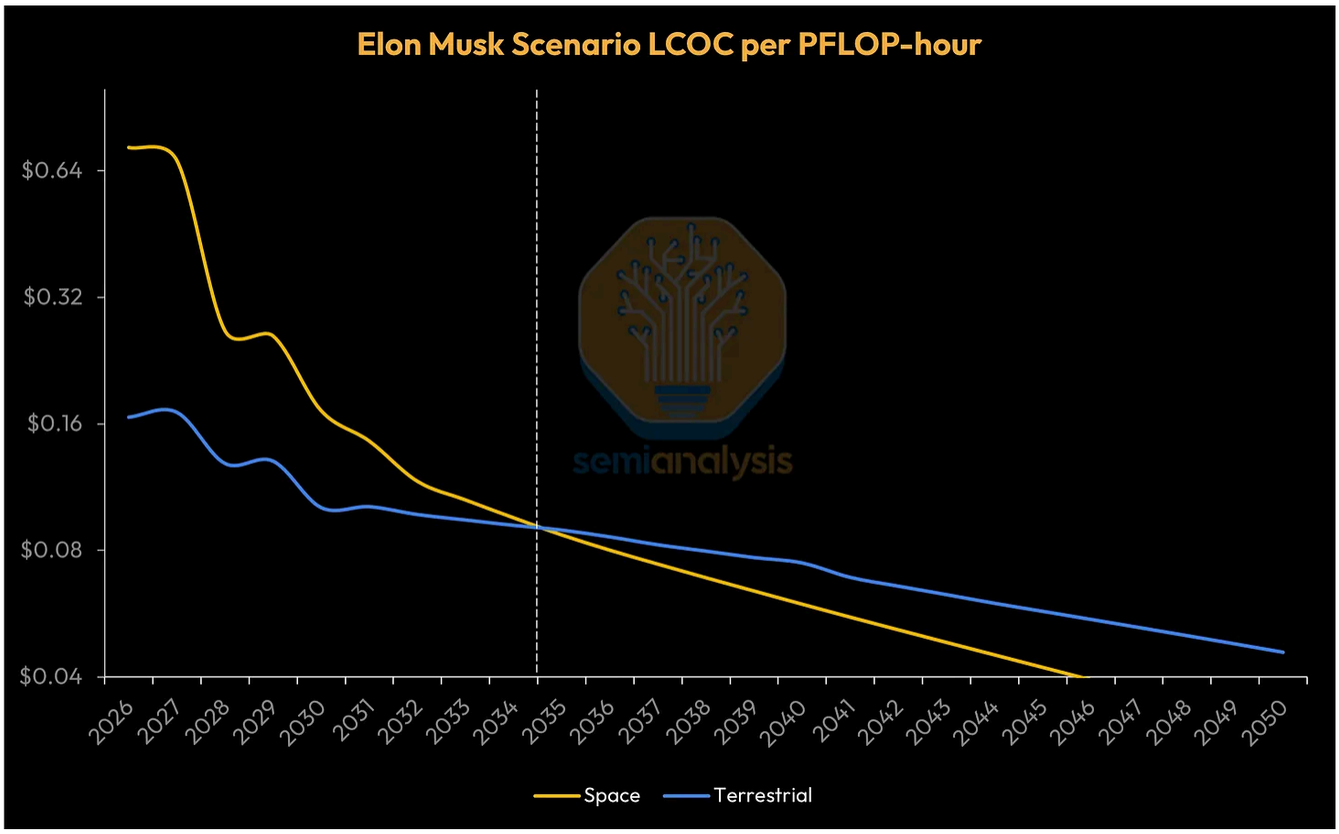
Source: SemiAnalysis AI Space Datacenter TCO Model

The bottom line is that overcoming the key fundamental obstacles to deploying space compute discussed above, as well as successfully scaling technology and launch costs is necessary to deploy space datacenters. It is the potential shortfall of terrestrial capacity that will drive actual demand rather than further cost differentials.

In the Elon Musk scenario, space cost improvements as well as higher terrestrial costs drive us to space-earth cost parity by ~2034, ~5 years earlier than in our base scenario.



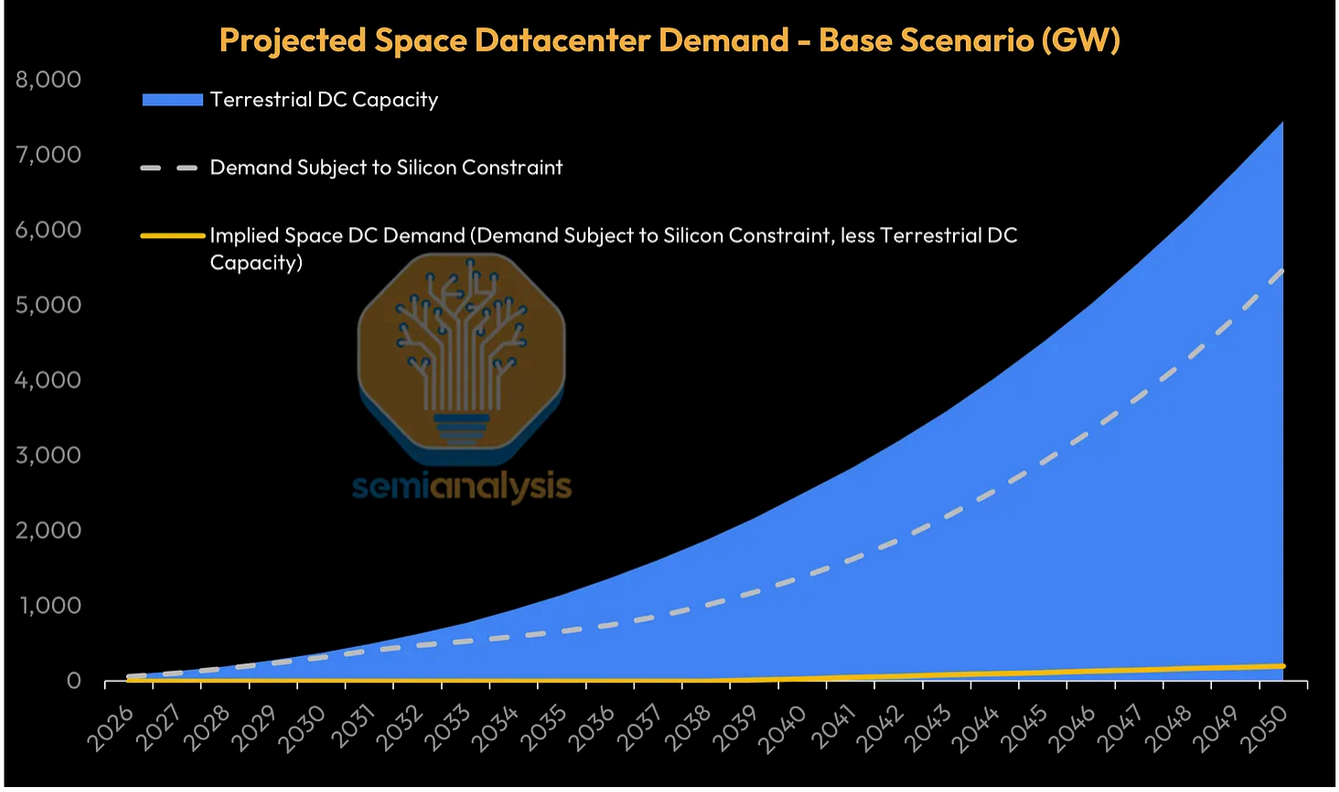
Source: SemiAnalysis AI Space Datacenter TCO Model



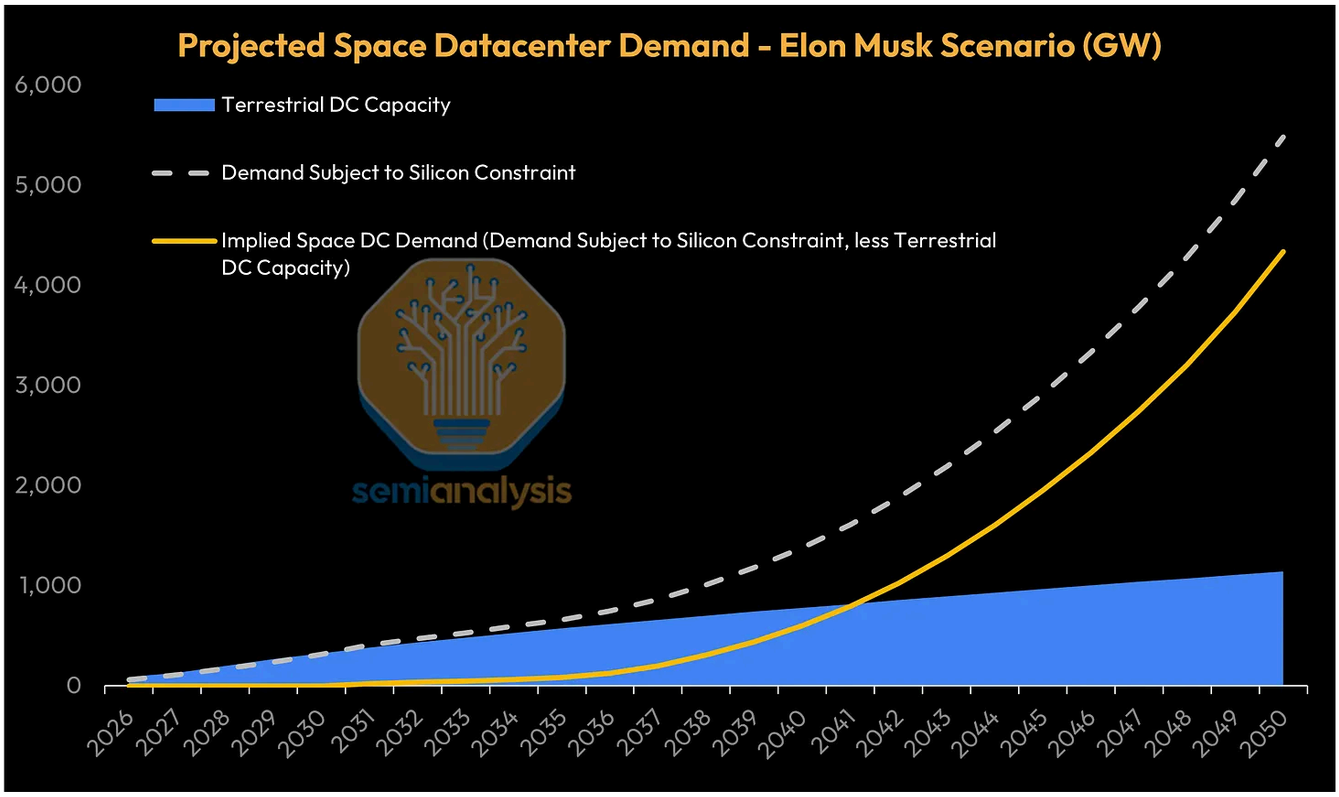
Source: SemiAnalysis AI Space Datacenter TCO Model

The Elon Musk scenario models a significantly lower terrestrial datacenter cumulative capacity - 576 GW by 2035 vs 1,150 GW in our base scenario, though proponents of large-scale space DC deployments undoubtedly believe this capacity number will land well below 300 GW.

Space-Earth cost parity is achieved in ~2034 in the Elon Musk scenario, opening the door to economic space deployments, just as critical IT power demand (subject to silicon constraints) decisively exceeds terrestrial capacity.

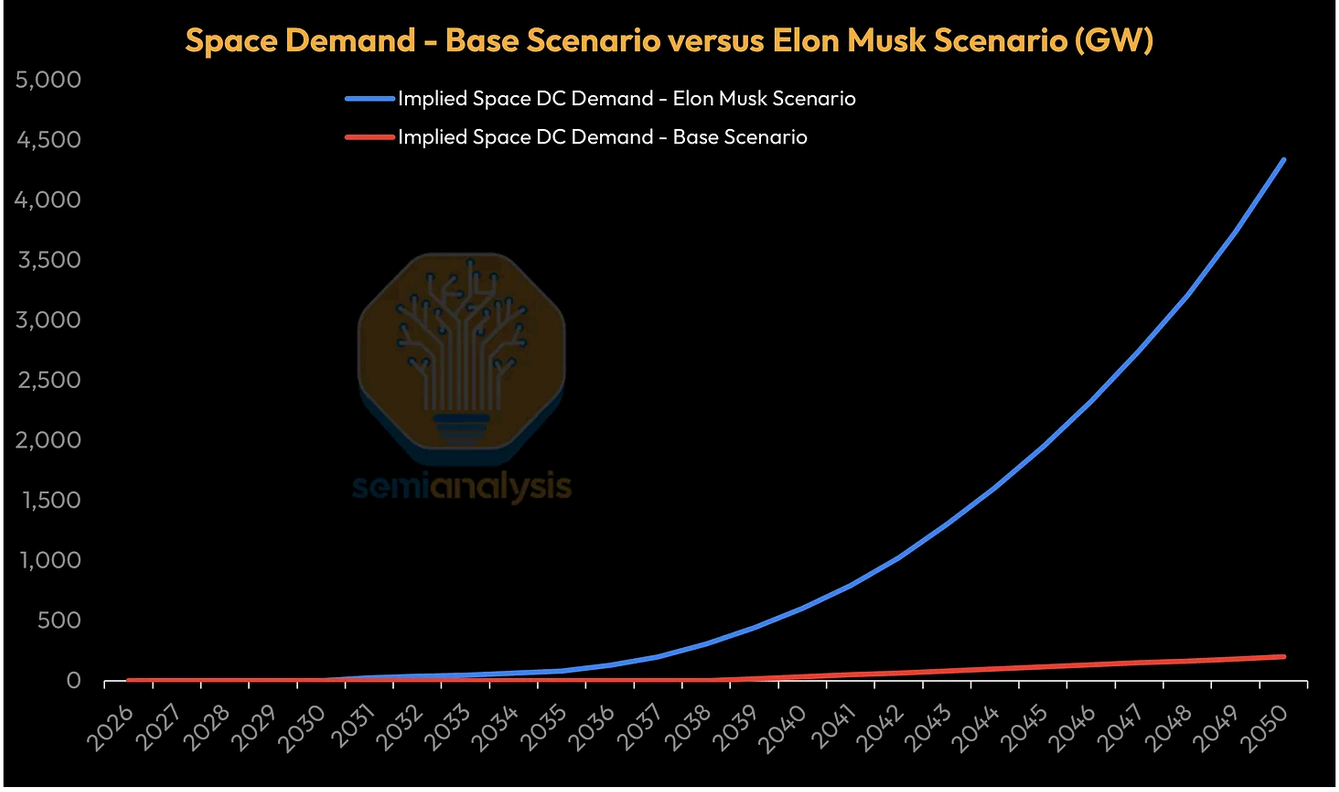


Source: SemiAnalysis AI Space Datacenter TCO Model



Source: SemiAnalysis AI Space Datacenter TCO Model

The result is a much earlier and much larger lift off in space demand. The below chart shows the install base of space datacenters in both scenarios.



Source: SemiAnalysis AI Space Datacenter TCO Model

As Musk has stated, an unexpected surge in end demand may be the factor that would decisively push compute into space. A demand scenario that underpins our Elon Musk scenario could look like this: 100W of AI compute per person globally, with every one of the eight billion people on Earth running a continuous, heavy AI workload simultaneously. This requires 800 GW of critical IT power by the late 2030s, 20-30x today's AI demand, implying north of $8 trillion in annual AI cloud revenue which is equivalent to roughly 28% of current US GDP generated by a single product category.

To understand when space datacenters make sense, one must consider SpaceX's vision: forecasting a TAM of $26.5T in AI ($22.7T enterprise applications, $2.4T infrastructure) - more aspirational than even our upside case. Consider also SpaceX's 100 TW-per-year compensation gate for Musk, 125 times larger than the 800 GW demand scenario, deployed via space datacenters in the long-term, underpinned by breakaway AI compute demand.

Simply put, there is not enough ground DC runway to realistically achieve any component of the vision above in a cost-effective and/or timely manner.

The economic incentives for operationalizing and optimizing (i.e. costs) a space datacenter program are high, and though critical engineering roadblocks remain, the economic opportunity driven by SpaceX will surely incentivize the supply chain and ecosystem for step-change improvements across the board.

Part Four: Space Datacenter Anatomy and Capital Costs

How Does a Space Datacenter Work? How is it Different from Terrestrial Datacenters?

A Space-based datacenter is radically different to a terrestrial datacenter. Like terrestrial datacenters, space datacenters have to physically house the compute, provide power input as well as heat dissipation.

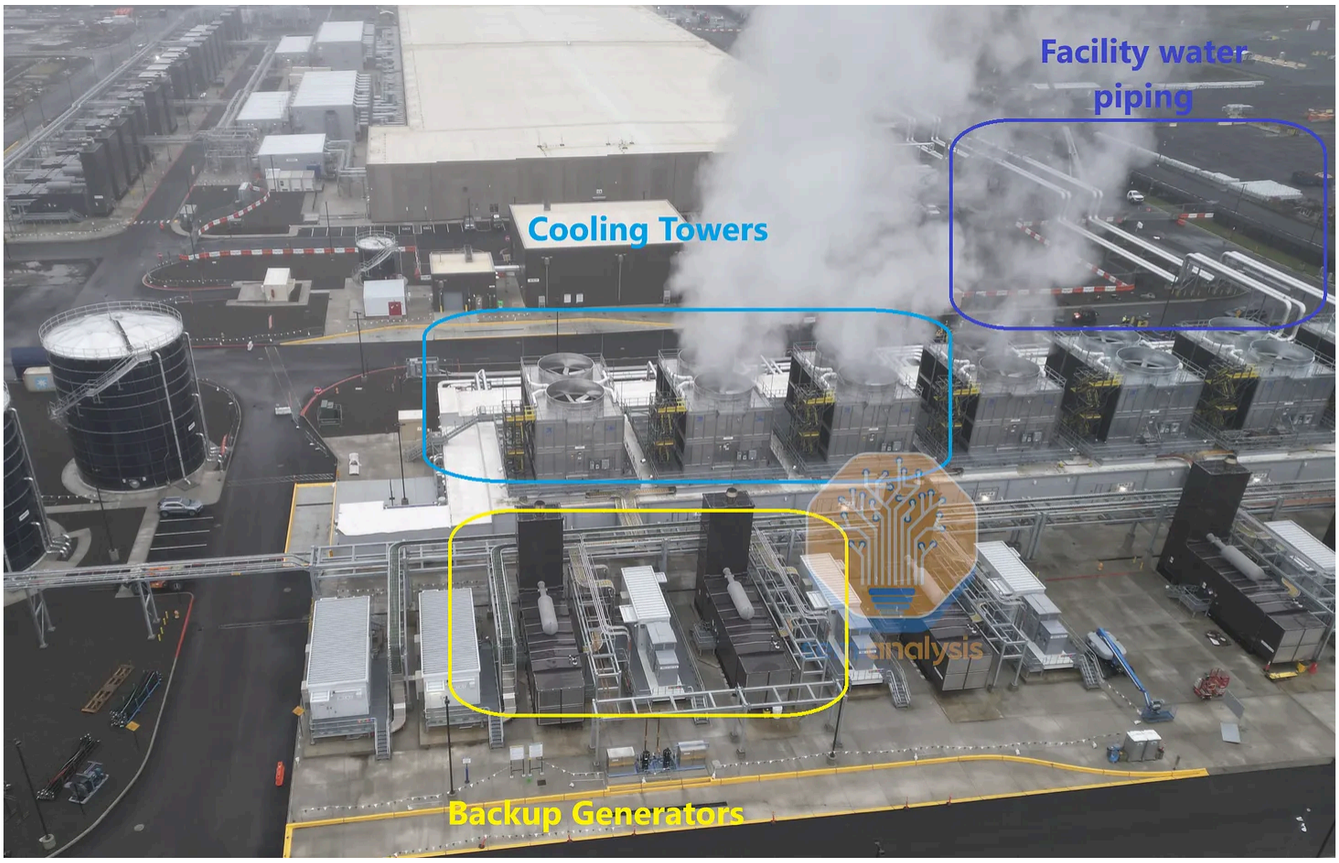
On the ground, heat dissipation translates to buildings, BTM energy or utility connections with backup generators, and industrial convection cooling. In orbit, we instead use spacecraft buses, solar arrays, and thermal radiators respectively.

We have built up a concept of what a space datacenter could be like based on engineering first principles. The final design of space datacenters could look different - but we want to illustrate the fundamental concepts at play. Our AI Space Datacenter TCO Model builds up every single space cost line item from first principles and allows users to modify these assumptions based on their own assessment of the pace of technological advancement and cost scaling.

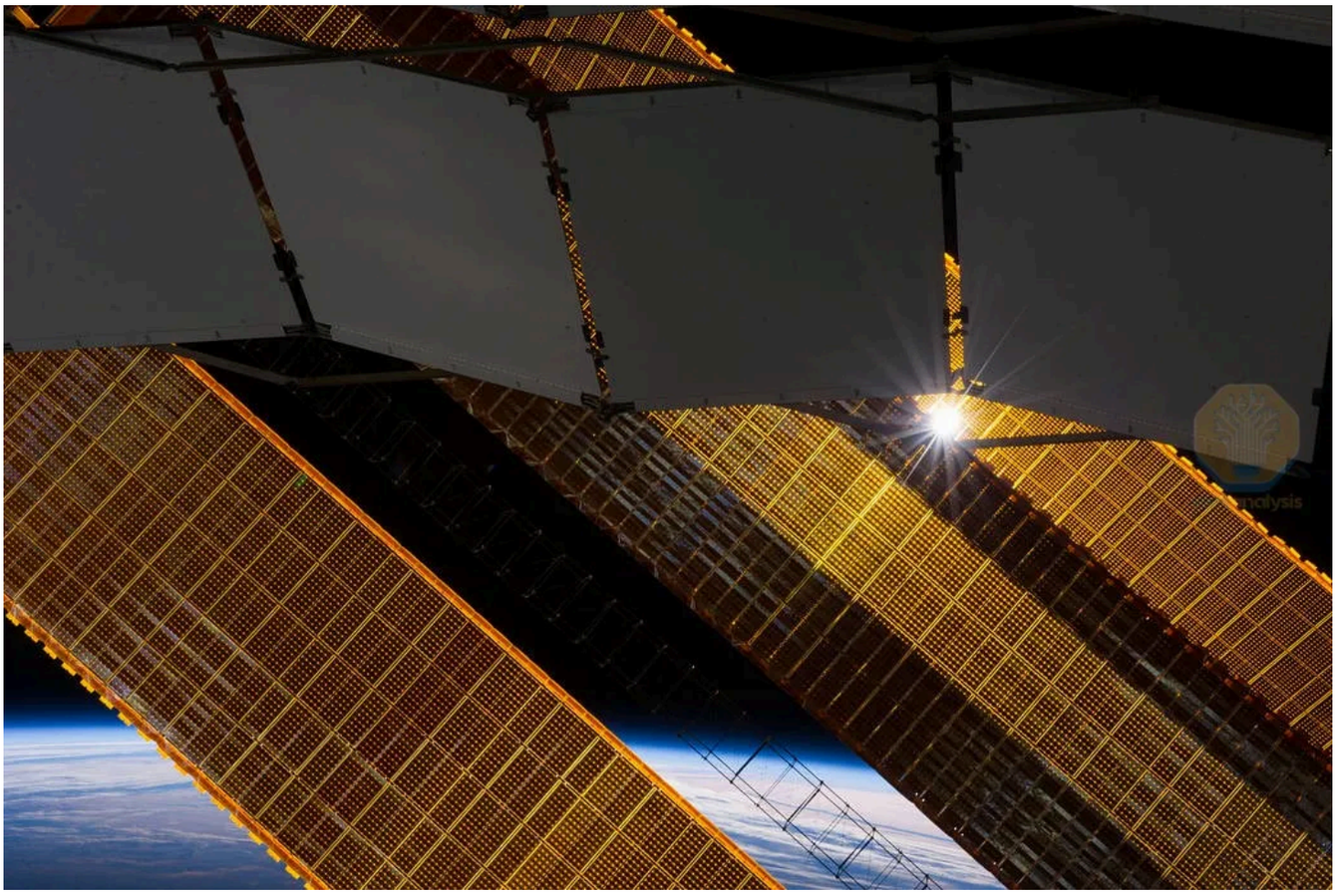
A space datacenter consists of two major groups of components - **Power and Thermal** as well as **Compute Payload and Facility**.

Compute Payload and Facility consist of the GPUs, CPUs, storage, networking to connect key compute and storage components among other IT equipment. **Power and Thermal** consists of the Solar Panels to provide power to the space datacenter, while radiators are responsible for removing heat from the datacenter.

For a terrestrial datacenter, power typically comes from grid power, while removing heat uses cooling towers and facility water. For a space datacenter, power will be provided by solar energy, and the only way to remove heat is through radiators.



Source: SemiAnalysis AI Datacenter Model



Source: NASA

Let's step through each of these key systems within a space datacenter:

Power: Solar Arrays

We explained earlier that not all orbits receive "free 24 hour solar energy". Most LEO orbits receive sunlight around ~60% of the time. We presented the retrograde Sun-Synchronous Orbit (SSO) as the ideal orbit to minimize (though not fully eliminate) battery requirements.

A detailed comparison of solar irradiance can be made for terrestrial, LEO, and SSO conditions, which will show that the physical performance advantage of space over terrestrial is meaningful. Solar irradiance at orbit altitude is 1,361 W/m^2 (the solar constant at top of atmosphere), compared to 1,000 W/m^2 Standard Test Conditions for terrestrial solar panels.

We further account for atmospheric attenuation of 27% at clear-sky conditions for terrestrial solar deployments, sunlit hours per day (explained earlier in the LEO and SSO section), and a weather discount factor (20-25% best-case for sunny regions, up to 40-50% as a global average) to result in the bottom-line effective solar irradiance in the table below.

Solar Irradiance in Terrestrial versus LEO vs Dawn-Dusk SSO

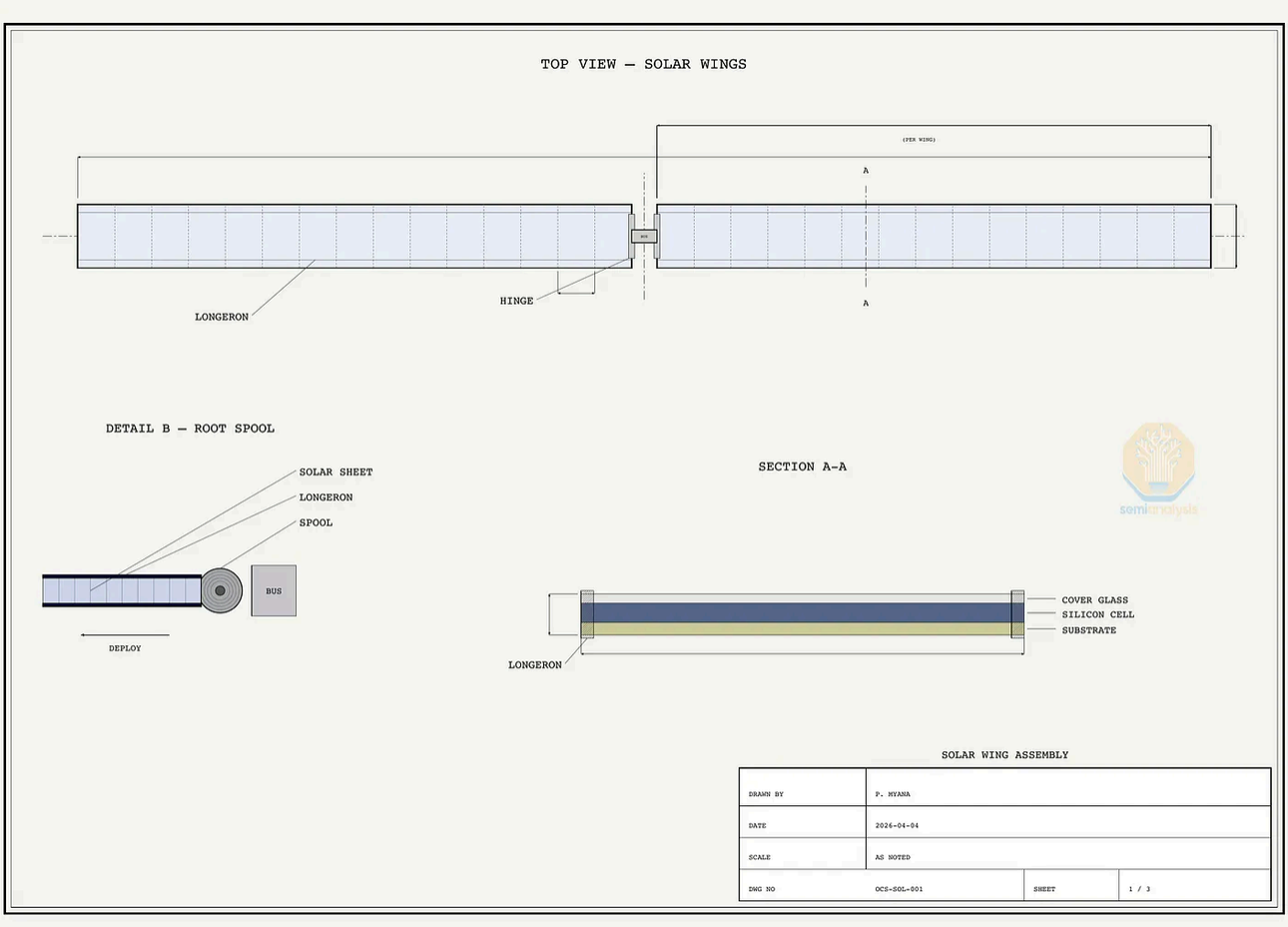
Parameter	Terrestrial	LEO	Dawn-Dusk SSO
Peak solar irradiance	1,000 W/m²	1,361 W/m²	1,361 W/m²
Atmospheric attenuation	27.0%	0.0%	0.0%
Attenuated peak irradiance	730 W/m²	1,361 W/m²	1,361 W/m²
Sunlit hours per day	12.0 hr	14.4 hr	23.5 hr
Weather/cloud discount factor[1]	25.0%	0.0%	0.0%
Effective solar irradiance	274 W/m²	817 W/m²	1,334 W/m²

1. Varies based on location and cloud cover. Best case 20-25% in sunny, low-cloud cover regions.

Source: SemiAnalysis

In a space datacenter setting with no electricity bill, capital costs replace operating costs. The solar array is sized to the total bus power requirement, which consists of the IT load and power needed for thermal management, attitude control, and communications overhead.

Below is a technical schematic of the solar wings used in our Space Datacenter concept. A 3-D interactive CAD model of this and other components can be accessed in our AI Space Datacenter TCO Model.



Source: Astrocompute, SemiAnalysis

Thermal: Active Cooling Loops and Radiator

Radiators are where terrestrial datacenters and space datacenters are similar, yet different. Any datacenter needs to remove heat from the IT devices. Terrestrial datacenters can sink heat into the atmosphere via convection and they do that by using cooling towers, chillers, and pumped water loops. Terrestrial datacenters expend

a lot of power doing this, consuming 25–40% of total facility power in doing so (reflected in PUE values of 1.25–1.40 respectively).

In space, unfortunately you're in a vacuum. That's why it's called "space". There is no air and therefore no convection. All waste heat must be rejected by thermal radiation, which is governed by the Stefan-Boltzmann law (the power radiated from a black body is proportionate to the fourth power of the temperature). At the operating temperatures typical of GPU cold plates at the upper range of a GPU's operating limit (~350 K or ~80°C), with a realistic LEO effective sink temperature of ~255 K, the radiator rejects approximately 880 W/m^2.

The baseline construction material used for the radiator is solid Aluminum (Al) 6061 - a 98% Aluminum, 1% Magnesium, 0.6% Silicon, 0.3% Copper and 0.2% Chromium alloy. While Falcon 9 uses Al-Li 2195 for primary tank structures and Starship uses 304L stainless steel, Al 6061 remains a widely-used commodity alloy in spacecraft secondary structures and panel applications. A radiator panel built to roughly 3 mm thickness with embedded heat pipes yields an areal density of about 8 kg/m^2, and costs scale with respect to the mass of the radiator.

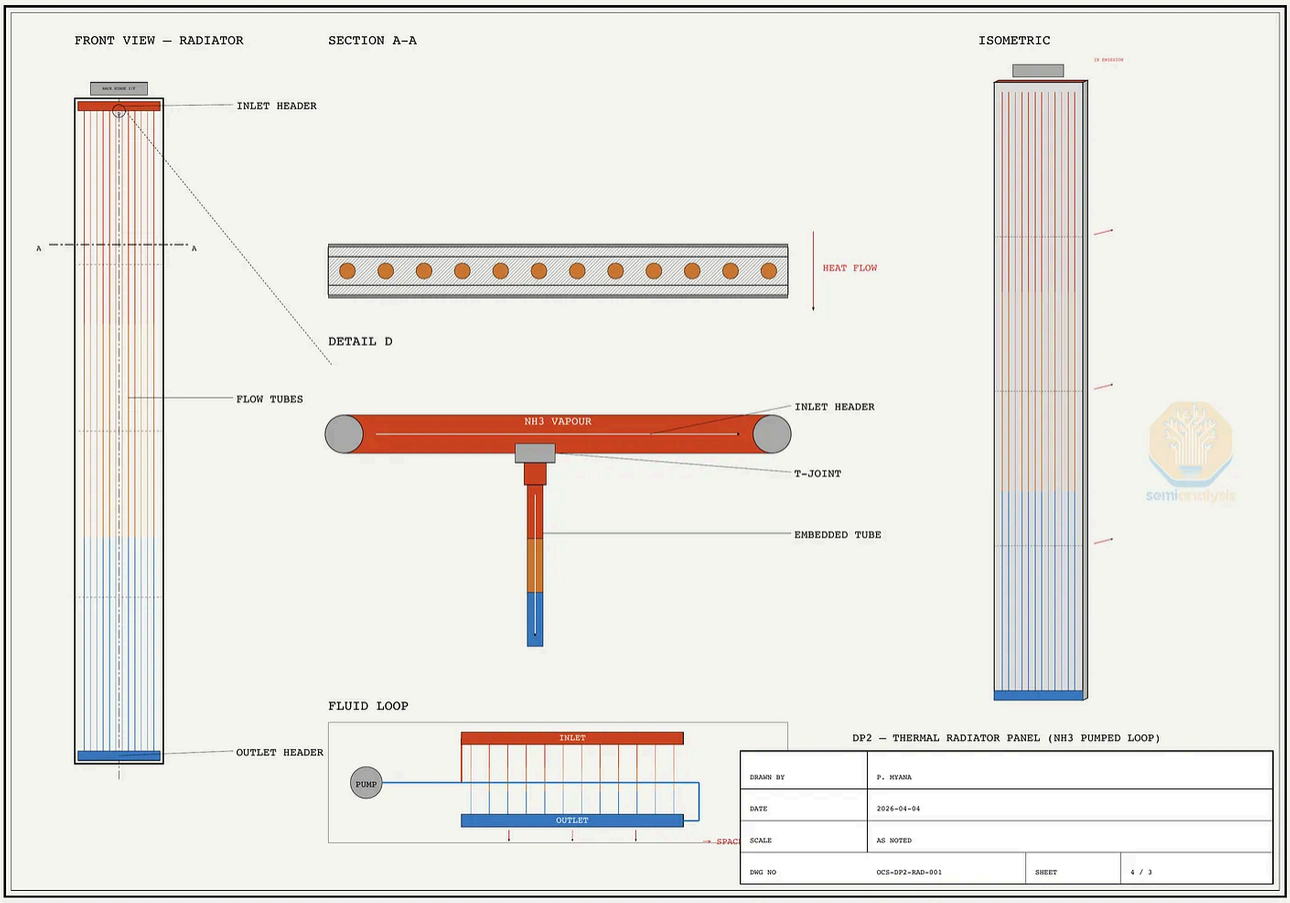
Each GPU couples to a cold plate, and a low-power pump circulates working fluid through manifold headers that carry waste heat from densely packed compute nodes out to deployable radiator panels. The fluid loop could be ammonia, CO_2, or a dielectric coolant architecture depending on operating temperature, toxicity tolerance, freeze risk, and serviceability.

This decouples the geometry of the compute payload from the geometry of the radiators. Compute can be packed into NVL8 (or NVL72) chassis with terrestrial-equivalent density, while radiators deploy as separate booms sized purely for radiative surface area. The heat path becomes: GPU → cold plate → coolant loop → distribution manifold → radiator panel → space. An active pump introduces a single point of failure and adds 2–4% to overall power consumption (analogous to terrestrial PUE overhead from chillers), but enables the system to scale far beyond what a direct-radiator-mount geometry would permit.

A research area that we project to gain significant interest in the coming years is the one of droplet radiators. These are radiators that pump the heat into a fluid of metals, spray it across space as droplets (which massively increases the surface area you are able to get compared to a flat panel), and then collects the droplets back just to repeat the process again. NASA has papers about droplet radiators from decades ago, but the interest in large platforms died after the end of the cold war and the decline of the space program at that time.

In our model, we include a section for radiators, with specific power (W/kg) constrained until droplet radiator breakthroughs in 2030. Post-droplet radiators, the specific energy of the satellites surges up. However, this advancement comes well after Starship's introduction and impacts our model well after. We expect power per chip to

continue to increase, which will drive up the power per mass of AI systems and make scaling of specific power important to good space datacenter economics.



Source: Astrocompute, SemiAnalysis

Compute Payload: The Facility

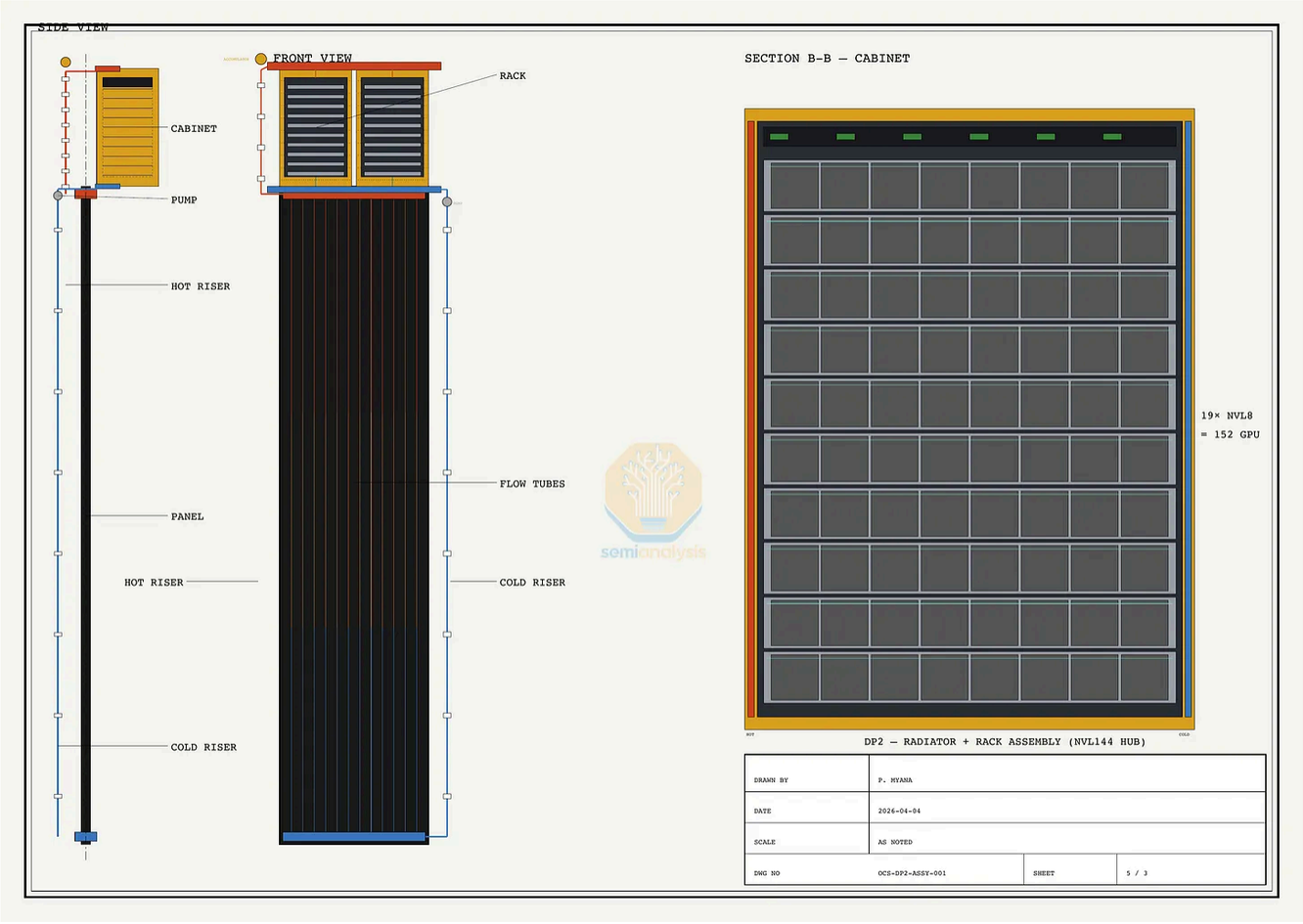
The orbital compute rack will not be a normal datacenter cabinet with fans removed. It will instead be a launch-rated, pressure-equalized aluminum structural module built around liquid-cooled compute trays. Each tray would carry an 8-GPU compute node and local NVSwitch/NVLink fabric, with short high-speed copper or optical links kept inside a tightly controlled mechanical envelope. Power will arrive from redundant spacecraft high-voltage DC buses, step down through fault-isolated DC/DC converter shelves, and be distributed through low-inductance copper busbars to the GPU trays.

The enclosure would be thermally selective rather than simply insulated. Non-radiating sun/albedo-facing surfaces would be covered with MLI blankets; radiator-coupled surfaces and thermal manifolds would remain exposed or conductively tied into the heat-rejection system. Heat would leave the GPUs through cold plates or vapor chambers into the pumped fluid loop mentioned in the prior section on radiators.

The rack would also need spacecraft-specific protection that a ground rack ignores: MMOD (orbital debris) shielding on ram-facing and high-risk surfaces, radiation shielding, latchup protection, ECC, scrubbing, watchdog resets, redundant power and pump paths, and launch-qualified structural mounts. It would not have perforated doors, blowers, hot/cold aisles, air filters, front service doors, or floor-based cable management. The compute fabric may resemble a terrestrial NVLink domain, but the

mechanical, thermal, power, radiation, and debris-protection envelope would be entirely different.

Networking uses high-bandwidth optical or RF links for data egress, with NVLink for intra-node GPU communication and Ethernet or InfiniBand for inter-node fabric. Orbital latency to ground stations is on the order of 5–20 ms for LEO, which is perfect for batch + inference.

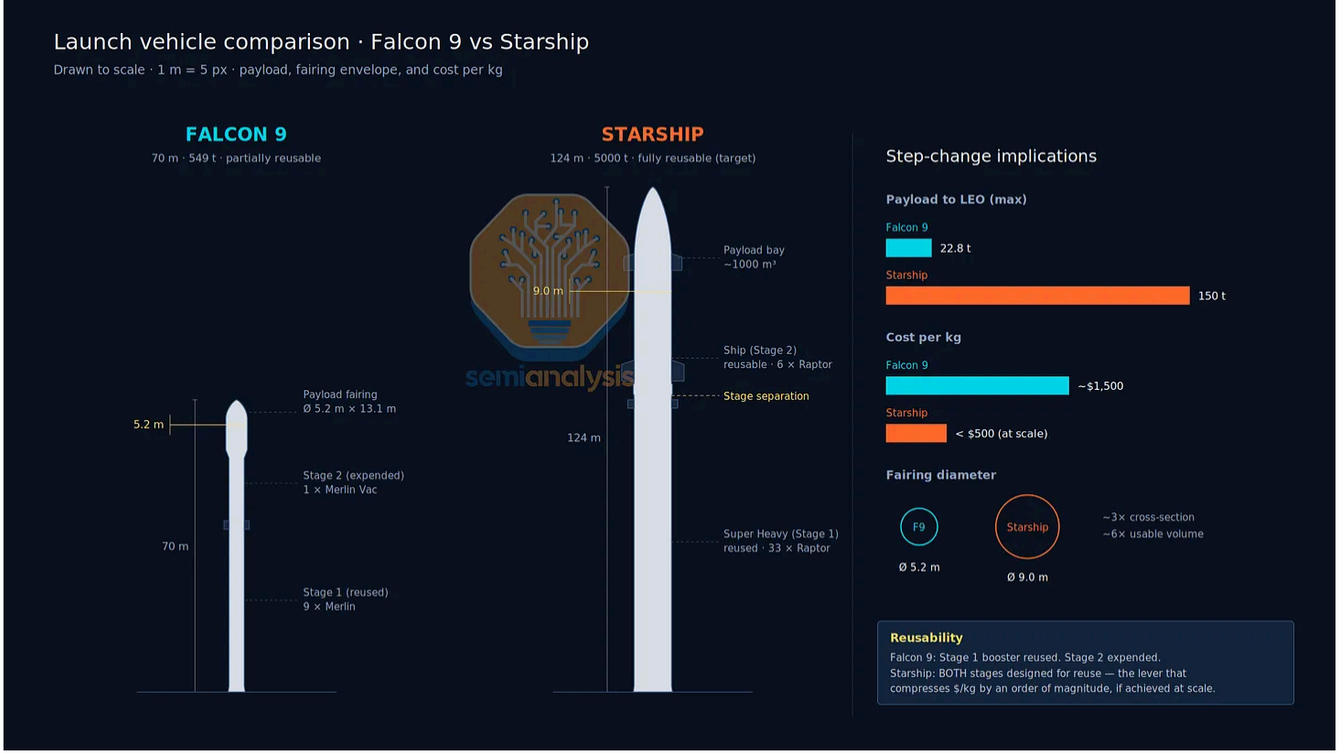


Source: Astrocompute, SemiAnalysis

Launch Vehicle: Getting to Orbit

The launch vehicle defines the mass and volume envelope for the entire system. Two vehicles matter today. Falcon 9, in expendable or rideshare configuration, delivers roughly 16,000–22,800 kg to LEO in a 5.2-meter fairing at $1,200-$1,700/kg (internal cost) depending on load. Starship targets 100–150 tons to LEO in a 9-meter fairing at ~$500/kg (internal cost), a step change that reshapes design tradeoffs. Earlier V2 flights will be in the range of ~50 tons. With Starship, you can use heavier, cheaper, more robust structures (thicker aluminum plates, standard extrusions) rather than the precision-folded, mass-optimized designs that Falcon 9's tighter envelope forces.

Both vehicles are two-stage. The first stage provides initial boost, then separates and returns to the launch pad for reuse. The second stage completes orbital insertion, circularizing at the target altitude (typically 500–800 km LEO). Starship's second stage (Ship) is itself designed for reuse and is the key innovation behind its projected $/kg compression. We cover the economic implications in detail in the section on launch costs.



Source: Astrocompute, SemiAnalysis

Space Capex Deep Dive

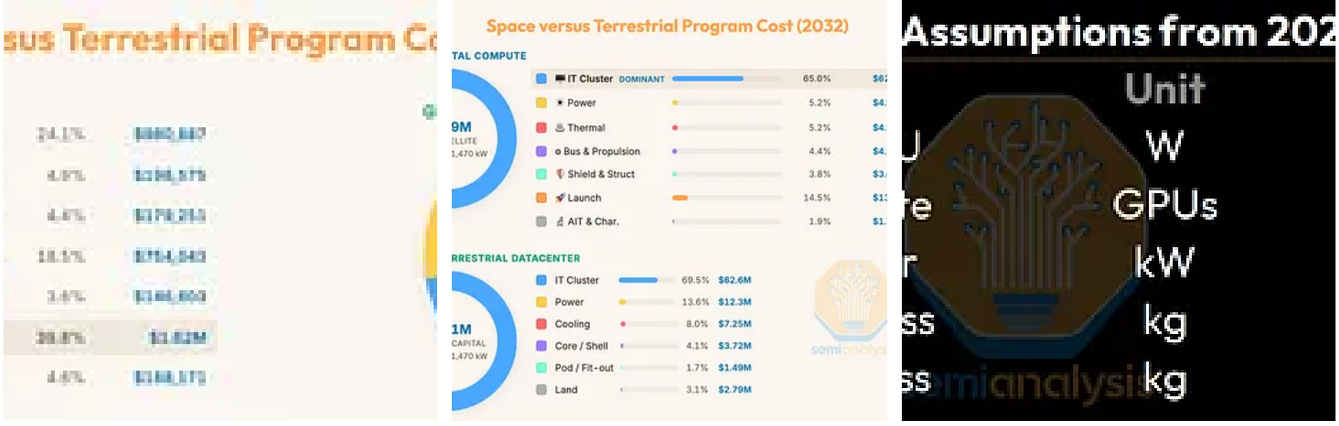
Our earlier section on our total cost of ownership framework briefly explained the cost structure for building terrestrial datacenters. As this is a space datacenter article, we will leave the discussion on terrestrial datacenter capex out for now. Those that are interested can read much more in our prior articles focused on the anatomy of a datacenter - we have one article focused on power and one on cooling.

Let's now dive into each category of capex for a space based datacenter and explain our cost estimate derivations in a good amount of detail. Building a space-based datacenter introduces another layer of complexity - not only do we have to build the facility and infrastructure for the datacenter itself, we have the added step of deploying these into space. This analysis will focus only on costs for our base case scenario.

The below interactive chart from the AI Space Datacenter TCO Model shows the key categories of space datacenter capex in 2026 and 2032.

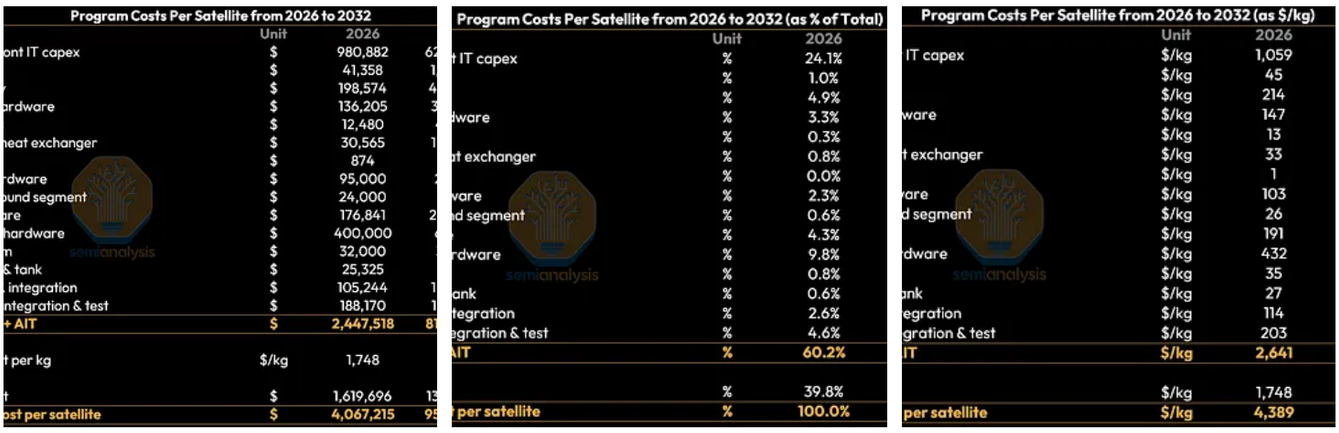
In the early years of space datacenters, costs will be dominated by launch costs (40% of program cost) as well as bus and propulsion costs (19% of program cost). IT Cluster costs (GPUs and other IT Equipment) will stand at only 24% of program cost.

As launch costs decline and as space datacenters become physically larger, scaling against bus and propulsion costs, these two major cost lines decrease as a percentage of program costs to only 15% for launch and 4% for bus and propulsion by 2032. Improvements in specific power of solar arrays and radiators also yield cost scaling. By 2032, IT Cluster program costs for Space Datacenters stand at 65% of program costs - a significant improvement from only 24% in 2026. In the long-run, most costs enjoy scale benefits as technology matures and we deploy larger space datacenters, but launch, power and thermal costs scale with IT cluster critical IT power.



Source: SemiAnalysis AI Space Datacenter TCO Model

The below tables outline Space program costs in absolute dollars, percent of total program cost, as well as per kg of total wet spacecraft mass.



Source: SemiAnalysis AI Space Datacenter TCO Model

USD per all-in Critical IT power is a useful measure of how much cost it takes to deploy one watt of IT power into space. If we analyze costs by looking at total cost per all-in Critical IT power in watts, total program cost falls from $132/W to only $65/W by 2032, with GPU hardware costs dominating the equation. We also see many cost items such as propulsion hardware, feed system, communication hardware among others scale considerably in just a few years as deploying 1MW+ datacenters by 2032 vs just a 30kW+ pilot datacenters allows these items to scale considerably. Advances in radiator and solar technology also help, but scale benefits will decelerate in coming years.

Program Costs Per Satellite from 2026 to 2032 (as $/W)			
	Unit	2026	2032
Space upfront IT capex	$/W	32.09	42.42
Shielding	$/W	1.35	1.16
Solar array	$/W	6.50	3.40
Radiator hardware	$/W	4.46	2.19
Cold plates	$/W	0.41	0.31
Pump and heat exchanger	$/W	1.00	0.89
Battery	$/W	0.03	0.01
Comms hardware	$/W	3.11	0.15
Comms ground segment	$/W	0.79	0.04
Bus hardware	$/W	5.79	1.98
Propulsion hardware	$/W	13.09	0.45
Feed system	$/W	1.05	0.21
Propellant & tank	$/W	0.83	0.03
Structure & integration	$/W	3.44	1.34
Assembly, integration & test	$/W	6.16	1.21
Hardware + AIT	**$/W**	**80.08**	**55.80**
Launch cost	$/W	52.99	9.44
Program cost per satellite	**$/W**	**133.07**	**65.24**

Source: SemiAnalysis AI Space Datacenter TCO Model

Now that we have a brief overview of the overall cost structure of space datacenters, let us examine each major cost item in detail.

Solar Array Cost

Sun Synchronous Orbit (SSO) will be the preferred orbit for space-based datacenters. While reaching SSO requires greater delta-V (change in velocity) than reaching Low Earth Orbit (LEO), the payoff is uninterrupted sunlight. As covered earlier, a dawn-dusk SSO solar array sees the sun continuously except for ~5% of the year, which is why every constellation wants access.

Continuous illumination also changes the panel itself. Without atmosphere, there is no need for protective glass, moisture-resistant materials, or IP67-rated junction boxes, and panels shed most of the weight ground arrays carry. Cell chemistry then becomes the tradeoff: standard silicon cells achieve 20–22% efficiency, while triple-junction gallium arsenide (GaAs) cells reach 28–32% but cost 10–20x more per watt. For 30 kW to 1 MW compute payloads, silicon's mass and cost advantage at Starlink-class production outweighs GaAs efficiency. At these volumes, solar panel costs can reach a cell cost of less than $0.30/W. Combining module, substrate and harness costs, we can achieve a total cost of $2.70/W. For our 30kW space datacenter deployed in 2026, this means a cost of $98,000 before taking deployment mechanisms into account.

At smaller volumes, the deployment mechanism (fold-out booms, hinges, latches) dominates panel cost, driven mainly by labor and NRE. For our 30kW space datacenter, this doubles the overall solar array cost to $198,500. As in-space manufacturing matures and arrays deploy as flat-packed rolls rather than articulated

panels, that cost should fall further. By 2032, for our ~1.5MW space datacenter the deployment mechanism is just $320k out of $5M total solar panel costs.

Taking into account other factors including annual degradation rate and BOL-EOL sizing (i.e. oversizing at beginning-of-life to account for end-of-life degradation), we see modest cost per watt improvements from $2.70/W for the solar panel in 2026 to $2.54/W in 2032. Considerable specific power improvements (greater W/kg) will drive down solar array mass, helping with launch costs.

Solar Array Cost Buildup

	Unit	2026	2032
Cell technology		Standard Si (20%)	3-jct perovskite
Cell efficiency (BOL)	%	20.0%	28.8%
Solar constant (AM0)	W/m²	1366	1366
Panel areal mass density	kg/m²	1.50	0.87
Power per m² (BOL)	W/m²	273	394
Panel Area	m²	111.88	3,729.38
Specific power (BOL)	W/kg	182.13	455.03
Cells cost per watt	$/W	0.30	0.28
Module assembly cost per watt	$/W	0.50	0.47
Substrate cost per watt	$/W	0.60	0.56
Harness cost per watt	$/W	1.30	1.22
Cost per watt	**$/W**	**2.70**	**2.54**
Deployment mechanism	$	100,000	321,834
Annual degradation rate	%/yr	2.5%	2.5%
Bus voltage	V	800	800
Housekeeping power overhead	%	6.0%	3.8%
Eclipse fraction (orbit-averaged)	%	5.0%	5.0%
BOL array power (sized for EOL)	kW	36.5	1838.5
Solar array mass	kg	200	4,040
Solar array cost	**$**	**198,574**	**4,992,365**
Total Cost per kg	$/kg	991	1,236
Total Cost per W of Critical IT Power	$/W	6.50	3.40

Source: SemiAnalysis AI Space Datacenter TCO Model

The array also has to absorb two known losses. It must power the satellite when lit and charge batteries for eclipse, and cells degrade about 2.5%/year, so a 5-year mission ends with roughly 12% less output than at launch. Designers oversize the array accordingly.

The bigger catch is overcrowding. ESA's space debris office tracks thousands of objects in SSO, with the count climbing 20% annually, and adding thousands of large-cross-section compute satellites compounds the risk. The real threat isn't the 34,000 trackable objects >10 cm; it's the estimated 130+ million sub-cm particles carrying grenade-level kinetic energy (7.5 km/s). A 1 mm aluminum particle at that velocity

delivers ~15 J/mm^2 on impact. Space datacenters are especially exposed because their large radiators for waste heat rejection are thin and thermally critical. Satellites must therefore be meaningfully oversized to absorb years of strikes from small, nearly undetectable debris.

Radiator Hardware Cost

Space datacenters reject waste heat exclusively through radiation. Without an atmosphere present, there's no convection to carry heat away — every watt of electrical power consumed by the satellite must eventually leave as infrared photons radiated into deep space. This places radiator design at the center of space datacenter engineering, and makes thermal management one of the most expensive subsystems in the entire capex stack.

A space thermal system has three distinct systems, each with its own mass and cost:

1. **Cold plates**: Cold plates attach directly to GPU modules and conduct heat through a thermal interface material into the radiator substrate. Mass: 0.7 kg per GPU; cost: $700–$800 per GPU for our B300 concept. $12,480 for the entire 16 GPU cluster. Cold plate mass scales sub-linearly with respect to chip TDP while we model cost per watt improving modestly over time.
2. **Pumps, Heat Exchanger, Transport Loop**: Pumps, accumulators, plumbing manifolds, working fluid, and fittings that move heat from cold plates to the radiators. Mass: 0.5–1.0 kg/kW; cost: $1,000/kg and decreasing. At 30 kW this is 15kg and $30k.
3. **Radiator panels**: Large flat panels coated in high-emissivity paint (Carbon Black), deployed on booms. Mass: 8 kg/m^2 for our 2026 space datacenter but materials advances drop this to only 3.5 kg/m^2 for our 2032 space datacenter; cost: $3,000–3,500/m^2 in 2026 dropping to $1,710/m^2 by 2032. Our 30 kW satellite deployed in 2026 requires a radiator assembly including deployment mechanism with a mass of ~390 kg costing ~$135k. By 2032, our ~1.5MW space datacenter will use a radiator with a mass of 7,500kg with a cost of $3.2M.

Starting with radiators, thermal energy radiated via photonics is governed by the Stefan-Boltzmann law, which we have referenced throughout the piece,

$$P_{rad} = \varepsilon \sigma A \ (T_{rad}^4 - T_{sink}^4)$$

where:

P_rad: radiated power (Watts). The heat the radiator is dumping into space.
ε: Emissivity. How efficiently the surface radiates compared to the perfect blackbody.
σ: Stefan-Boltzmann constant
A: Radiator area (m^2). The surface area of the radiator that emits the heat.
T_rad: radiator temperature
T_sink: sink temperature (the cold environment the radiator is dumping the heat into)

At a radiator temperature of 343 K and a sink temperature of 255 K (effective sink including Earth IR and albedo), emissivity 0.95, a double sided radiator rejects approximately 880 W/m^2. A 30 kW satellite therefore needs roughly 42 m^2 of radiator area to reject waste heat.

The fourth-power scaling in the Stefan-Boltzmann equation has a critical implication for design: running the radiator hot is dramatically more efficient than running it cool. A radiator at 370K rejects 2.3x more heat per square meter than one at 320K. This drives the central design imperative for space datacenters — run electronics as hot as they will reliably tolerate, so that the radiator can also operate hot, minimizing the radiator area, mass, and cost required to reject a given thermal load. The hotter we run the chip, the less radiator area is needed, and the better costs scale. The limit comes from electronics reliability, not thermal physics — chip temperatures above ~85-90°C accelerate failure modes and reduce mean time to failure.

The modern material of choice is Aluminum 6061-T6 flat plate. It is the most weldable heat-treatable aluminum alloy, and it is possible to work with this material to friction stir weld large tank barrels and structural joints at high production rates. Friction Stir Welding can be done with automated tooling, which enables high rate manufacturing that defines SpaceX's cost advantage. It's the most commonly produced aluminum alloy and a commodity, which is part of the supply chain leverage that SpaceX exploits (similarly to how it uses silicon solar instead of Gallium Arsenide).

Aluminum 6061-T6 is a great candidate for radiator panels, as its isotropic thermal conductivity of 167 W/m·K provides great heat spreading across the radiating surface, eliminating the need for honeycomb sandwich construction that would otherwise require very skilled layup and autoclave bonding. One disadvantage of foregoing honeycomb construction is that without it you need ~33% more mass per square meter versus with honeycomb construction (8 vs 6 kg/m^2). However, at Starship's ultimate launch cost targets of $200–600/kg, the extra mass penalty is dwarfed by the savings from fabrication, as honeycomb processing can add $800+/$m^2$, increasing overall radiator cost by $32,000–150,000 for a 40 m^2 array — far exceeding the $16,000–84,000 launch penalty from the extra 80 kg of panel mass at Starship rates. This aligns with SpaceX's philosophy: use more of a cheap, simple material rather than less of an expensive, complex one.

Modern radiators cost on the order of thousands of dollars per kW for the panel alone, but in our model, we argue that costs can be even lower thanks to the use of Al 6061-T6, driving total cost per kW from $4,500/kW in 2026 down to only $2,000 per kW by 2032. With advances in material science, we think this cost can drop even further to below $2,000/kW by the 2030s — a 65% overall reduction in cost per W driven by two specific improvements. First, a doubling in effective heat rejection per m^2 from 880 W/m^2 to almost 1,400 W/m^2 through higher operating temperatures and improved radiator coatings. Second, a halving in areal density from 8 kg/m^2 down to 3.5 kg/m^2

through manufacturing improvements and potentially droplet radiator designs that pump liquid metal directly into space rather than radiating through panel surfaces.

Mass and cost for radiators scale linearly with power. Our 2026 space datacenter calculations suggest a specific power of ~80 W/kg, but innovations such as droplet radiators or the ability to increase operating temperature have great effects on specific power - our 2032 space datacenter design assumes radiator designs that can reach a specific power of 195 W/kg!

Specific power scales less dramatically beyond this point - reaching 300 W/kg as we approach 2050. The most dramatic gains are from now until the mid 2030s, but efficiency gains slow as the Stefan-Boltzmann law puts fundamental constraints on the effective heat rejection per m^2. The "run hot" lever only goes so far — chip temperature reliability limits constrain how aggressive radiator operating temperatures can be, and the fourth-power scaling that helps so much going from 320K to 370K becomes diminishing returns going from 370K to higher temperatures.

Our 2032 configuration also benefits from being launched on Starship, with Starship's larger diameter relaxing the volume constraint that forces complex radiator folding mechanisms. To factor in the deployment costs, we assume that 15% of the mass is allocated towards deployment structures, though this scales over time.

Radiator Hardware Cost Buildup

	Unit	2026	2032
Waste Heat to Reject	kW	32	1,526
Radiator Temperature	K	343	346
Sink Temperature	K	255	255
Emissivity	-	0.80	0.81
Thermal radiation emitted	W/m	628	660
Less: Earth infrared heat absorbed	W/m	19	20
Less: Reflected sunlight absorbed	W/m	8	8
Less: Direct sunlight absorbed	W/m	14	14
Net cooling per radiator face	W/m	587	619
Effective heat rejection per m²	**W/m²**	**881**	**928**
Required radiator + deployment area	m²	42	1,882
Panel areal density	kg/m²	8	4
Deployment / structure mass fraction	%	15%	14%
Cost per m² (panel)	$/m²	3,220	1,710
Radiator mass	kg	389	7,534
Radiator cost	**$**	**136,205**	**3,217,389**
Total Cost per kg	$/kg	350	427
Total Cost per W of Critical IT Power	$/W	4.46	2.19

Source: SemiAnalysis AI Space Datacenter TCO Model

How high historical radiator costs have distorted space datacenter cost analysis

A cursory analysis might just look at the cost of historical orbital radiator systems and conclude that space thermal economics are unworkable. The International Space Station's Active Thermal Control System (ATCS) is the most relevant large-scale reference for orbital heat rejection, and it consists of two parallel subsystems serving different parts of the station.

The External Active Thermal Control System (EATCS) is the main 70 kW heat rejection system for the crewed modules. Its Heat Rejection System (HRS) radiators deploy 48 panels arranged as 6 Orbital Replacement Units (ORUs) of 8 panels each, with each panel measuring 3.33 × 2.64 m (~8.8 m^2 per panel) for a total radiating area of 422 m^2 one-face (844 m^2 counting both radiating faces). The EATCS uses single-phase liquid ammonia circulated through Inconel 718 flow tubes bonded between aluminum extrusions for MMOD protection, with the radiating face sheets coated in Z-93 white silicate paint ($\varepsilon \approx 0.92$, $\alpha_s \approx 0.16$) to maximize heat rejection while minimizing solar absorption — a constraint imposed by ISS radiators being directly sun-illuminated. The sun-shielded radiator geometry described earlier for the compute platform avoids this constraint and enables use of higher-emissivity carbon-black coatings ($\varepsilon \approx 0.95$) without solar absorption penalty. The total EATCS cost is estimated at $340–500M, with Boeing as system prime and Lockheed Martin (Grand Prairie, TX) as radiator subcontractor.

The Photovoltaic Thermal Control System (PVTCS) is a separate, smaller heat rejection system cooling the solar array power electronics (batteries, DC-to-DC converters) on each of the four power modules (P4, S4, P6, S6). PVTCS uses 4 Photovoltaic Radiator (PVR) ORUs (28 panels total, ~170 m^2 one-face) and 8 active Pump Flow Control Subassemblies (PFCS) — two per power module — to reject up to 56 kW peak (14 kW per PVR) into space.

Combined, the U.S. segment ATCS represents approximately 126 kW peak rejection capacity (70 kW EATCS + 56 kW PVTCS) on separate ammonia loops, at an estimated total program cost of $570–830M, or roughly $4.5–6.6M per kW rejected. This cost breaks down approximately as follows:

International Space Station (ISS) Active Thermal Control System (ATCS) Costs			
Component	Estimated cost	Quantity / size	Unit economics
External Active Thermal Control System (EATCS)			
HRS deployable radiator panels	$140–200M	48 panels across 6 ORUs; 8.8 m² per panel	~$2.9–4.2M per panel; ~$332k–$474K/m² (one-face)
Pump modules (ammonia)	$80–120M	8 pump modules (4 active + 4 spare)	~$10–15M per module
Plumbing, accumulators, N2 pressurant	$40–60M	100m+ of tubing	Largely labor & EVA installation
Development / NRE	$80–120M	One-time	Boeing (system prime)
EATCS subtotal	**$340–500M**	**Rejects ~70 kW**	**~$4.9–7.1M per kW rejected**
Photovoltaic Thermal Control System (PVTCS)			
PV radiator deployable panels	$100–140M	28 panels across 4 PVR ORUs; ~6.0 m² per panel	~$3.6–5.0M per panel; ~$590k–$820k/m² (one-face)
Pump flow control subassemblies (PFCS)	$80–130M	8 PFCS units (2 per power module × 4 modules)	~$10–16M per PFCS
Plumbing & integration	$30–50M	Per-module ammonia loop	Allocated across 4 power modules
Development / NRE	$20–40M	One-time	Shared heritage with EATCS
PVTCS subtotal	**$230–330M**	**Rejects ~56 kW peak (14 kW per PVR)**	**~$4.1–5.9M per kW rejected**
Total ATCS (EATCS + PVTCS)	**$570–830M**	**~126 kW peak rejection**	**~$4.5–6.6M per kW rejected**

Source: SemiAnalysis

At $4.5–6.6M per kW of rejected heat, the ISS thermal system is almost 1,000× more expensive per watt than what we estimate for our 2026 30kW satellite. The high costs paid for the ISS deployment are not fundamental to the science or technology at play here. The ISS was designed in the late 1980s, built through the 1990s on cost-plus contracts, with every component individually qualified to NASA Class A standards, installed via Extravehicular Activity (EVA), i.e. space walks by astronauts at $100K+ per crew-hour, and produced in quantities of one at a time. The ISS numbers reflect 1980s-1990s design philosophy, not the cost floor for orbital thermal management.

A modern commercial satellite using deployable radiator panels from vendors like Maxar (FlexRad) or Redwire, with commercial-grade qualification and production tooling, achieves $1,500–2,000/m^2 for the radiator panels versus the ISS's implied $430,000/m^2. This is a nearly 300× cost reduction driven by volume production, commercial practices, and modern manufacturing. The ISS number is a reminder of how expensive thermal management can become and as such, it cannot be a helpful datapoint for forecasting space datacenter costs.

Cold Plates, Pump and Heat Exchanger

While our design distributes the GPUs across the radiator itself, we still need cold plates to remove heat from the chip, as well as an active cooling loop, pumps and heat exchangers to distribute the heat evenly across the radiator. For our ~1.5MW 2032 space datacenter concept, this means a cost of $450k for the cold plates, and $1.3M for the pumps, heat exchanger, manifold and fluid. The radiator panels themselves become cheaper over time, but the broader thermal system retains meaningful cost because the transport infrastructure (pumps, manifolds, working fluid) don't benefit from the same physics-driven cost reductions that radiator panels enjoy.

Cold Plate Cost Buildup			
	Unit	2026	2032
Cold plate mass per GPU	kg	0.70	1.26
TDP of Chip	W	1,200	3,876
Cost dollars per Watt	$/W	0.65	0.36
Cold plate cost per GPU	$/GPU	780	1,413
Total cold plate mass	kg	11	403
Total cold plate cost	**$**	**12,480**	**452,236**
Total Cost per kg	$/kg	1,114	1,123
Total Cost per W of Critical IT Power	$/W	0.41	0.31

Source: SemiAnalysis AI Space Datacenter TCO Model

Pump and Heat Exchanger Cost Buildup			
	Unit	2026	2032
Critical IT Power	W	30,565	1,469,528
Pump and heat exchanger cost per watt	$/W	1.00	0.89
Pump and heat exchanger mass per watt	kg/W	0.0005	0.0005
Pump and heat exchanger mass	kg	15	713
Total pump and heat exchanger cost	**$**	**30,565**	**1,301,770**
Total Cost per kg	$/kg	2,000	1,826
Total Cost per W of Critical IT Power	$/W	1.00	0.89

Source: SemiAnalysis AI Space Datacenter TCO Model

Bus Hardware: Attitude Determination and Control (ADCS), Onboard Computer (OBC), Power Management and Distribution (PMAD)

Bus Hardware consists of sensors, actuators, onboard computers and the power management and distribution systems needed to supply power to the IT payload.

The satellite must maintain its solar arrays pointed at the Sun and its antennas pointed at the ground simultaneously. ADCS achieves this through a combination of sensors (star trackers, inertial measurement units, magnetometers) and actuators (reaction wheels, magnetorquers, thrusters for large maneuvers).

There are a few main components covered by our ADCS and OBC estimates:

1. **Star trackers**
   Star trackers ($50K/unit, 1–2 units) measure attitude to <5 arcsecond accuracy by comparing observed star patterns to a catalogue.
   Flight heritage: decades. Mass: ~0.5–1 kg each.
2. **Reaction wheels**
   Reaction wheels ($200K/unit, 4 units for full 3-axis + redundancy): spinning flywheels that store angular momentum. Torquing the wheels changes satellite

attitude. These must be periodically desaturated using magnetorquers or thrusters when they reach speed limits.

3. **Magnetorquers** — Magnetorquers ($10K/axis, 3 axes): electromagnetic coils that interact with Earth's magnetic field to dump the momentum of the wheels. It cannot operate outside Earth's magnetosphere.
4. **Onboard Computer and Sensors** - $70-100k, Mass of 15-20kg.

Put together, for our 2032 ~1.5MW space datacenter, we estimate a cost of $735k for ADCS components, and an additional $106k for Onboard Computers and sensors with a combined mass of 125kg.

Turning to **Power Management and Distribution** (**PMAD**), every watt of solar power must be routed from array to battery to payload through copper (or occasionally aluminum) wiring. Harness mass scales with bus power divided by bus voltage: higher voltage means lower current for the same power, meaning thinner wire and lower harness mass. Our model uses 800V as a baseline.

Unfortunately power distribution mass scales proportionately to total datacenter critical IT power, though we do see some cost scaling. This means that power management and distribution is one of the largest components of mass for the space datacenter in later years in our forecast.

Bus Hardware Cost Buildup			
	Unit	2026	2032
ADCS actuator mass	kg	8	104
ADCS actuator cost per kg	$/kg	7,500	7,061
ADCS actuator cost	**$**	**60,000**	**734,637**
OBC and sensors mass	kg	13	20
OBC and sensors cost per kg	$/kg	5,358	5,199
OBC + sensors cost	**$**	**70,994**	**106,109**
PMAD cost per Watt	$/W	1.50	1.41
PMAD mass per Kilowatt	kg/kW	1.50	1.50
PMAD mass	kg	46	2,204
PMAD cost	**$**	**45,847**	**2,075,298**
Bus hardware total cost	**$**	**176,841**	**2,916,044**
Bus hardware total mass	kg	67	2,329
Total Cost per kg	$/kg	2,636	1,252
Total Cost per W of Critical IT Power	$/W	5.79	1.98

Source: SemiAnalysis AI Space Datacenter TCO Model

Propulsion Hardware, Feed System, Propellant and Tank Cost

Propulsion serves three purposes: orbit insertion (if the launch vehicle drops the satellite at the wrong altitude), station-keeping (resisting atmospheric drag), and end-

of-life deorbit. The dominant delta-V cost is deorbit at end of life: dropping from 650 km to 200 km where atmospheric drag will naturally complete reentry takes approximately 120–150 m/s. A Hall-effect thruster consuming Xenon or Argon handles this comfortably, but there is a minimum number of thrusters needed (two thrusters needed), making propulsion hardware costs very sub-scale for a small satellite like our 2026 30kW concept at $400,000 total.

We stick to two thrusters even with our 2032 concept, meaning costs are very similar, with the $354k estimated in that year benefiting from cost downs over 6 years. Instead, our feed system and propellant used will scale with the overall space datacenter mass, becoming a much larger percentage of costs and mass in later years.

Propulsion Hardware, Feed System, Propellant and Tank Cost Buildup

	Unit	2026	2032
Dry mass, excluding propulsion	kg	769	18,097
Drag compensation ΔV	m/s	50	50
Mission lifetime	year	5	10
De-orbit ΔV	m/s	150	150
Total ΔV budget	**m/s**	**400**	**650**
Number of thrusters	#	2	2
Thrust per thruster	N per m/s	0.08	0.08
Mass per thruster	kg	10	10
Total thruster hardware cost	**$**	**400,000**	**354,337**
Total thruster mass	kg	20	20
Total Cost per kg	$/kg	20,000	17,717
Total Cost per W of Critical IT Power	$/W	13.09	0.24
Feed system mass	kg	8	93
Feed system cost per kg	$ / kg	4,000	3,332
Total feed system cost	**$**	**32,000**	**308,212**
Total Cost per kg	$/kg	4,000	3,332
Total Cost per W of Critical IT Power	$/W	1.05	0.21
Specific impulse	s	1,500	1,500
Propellant and tank mass	kg	25	925
Propellant cost per kg	$ / kg	11	11
Tank cost	$	25,049	26,850
Total propellant and tank cost	**$**	**25,325**	**37,210**
Total Cost per kg	$/kg	1,028	40
Total Cost per W of Critical IT Power	$/W	0.83	0.03

Source: SemiAnalysis AI Space Datacenter TCO Model

Structure and Integration

Structure and Integration refers to structural elements of the space datacenters - the

physical hardware that holds all the systems together. This could include items such as the primary truss, various brackets, deployable mechanisms, mechanical harnesses among other items. It also includes the various labor and assembly costs needed to assemble the components of the structure. We model this at a constant 15% of the structural dry mass of the space datacenter, with a structure and integration rate starting at $1,000/kg, but with brisk cost downs over the years.

Structure and Integration Cost Buildup			
	Unit	2026	2032
Strucutral % of dry mass	kg	15%	15%
Structural mass	kg	105	2,365
Structure and integration cost / kg	$/kg	1,000	833
Structure & integration Cost	**$**	**105,244**	**1,970,184**

Source: SemiAnalysis AI Space Datacenter TCO Model

Assembly, Integration and Test (AI&T)

AI&T is the process of taking individually procured and tested components (solar panels, radiator panels, GPU modules, ADCS units, propulsion system) and physically assembling them into a working satellite, then verifying that the integrated system meets its performance specifications before launch. It is almost entirely a labor and facility cost. AI&T does not appear in a bill of materials because it produces no physical parts, but it is consistently one of the largest line items in a satellite program.

The 10% Rule - A widely-used industry heuristic is that AI&T costs approximately 10% of total hardware procurement. This includes: mechanical integration (bolting, bonding, cable routing, connector mate/demate), electrical integration (harness checkout, power-on testing, interface verification), functional testing (subsystem-level and system-level test sequences), environmental testing (vibration, acoustic, thermal vacuum, EMI/EMC), launch site operations (transportation, final integration, launch campaign), Environmental testing (particularly thermal vacuum testing, where the satellite is placed in a large vacuum chamber and cycled through the temperature range it will experience in orbit) is the most expensive and time-consuming test.

AI&T costs scale with volume production, and as such we start with AI&T costs of 15% of all non-GPU costs, with this ratio reaching 10% by 2032.

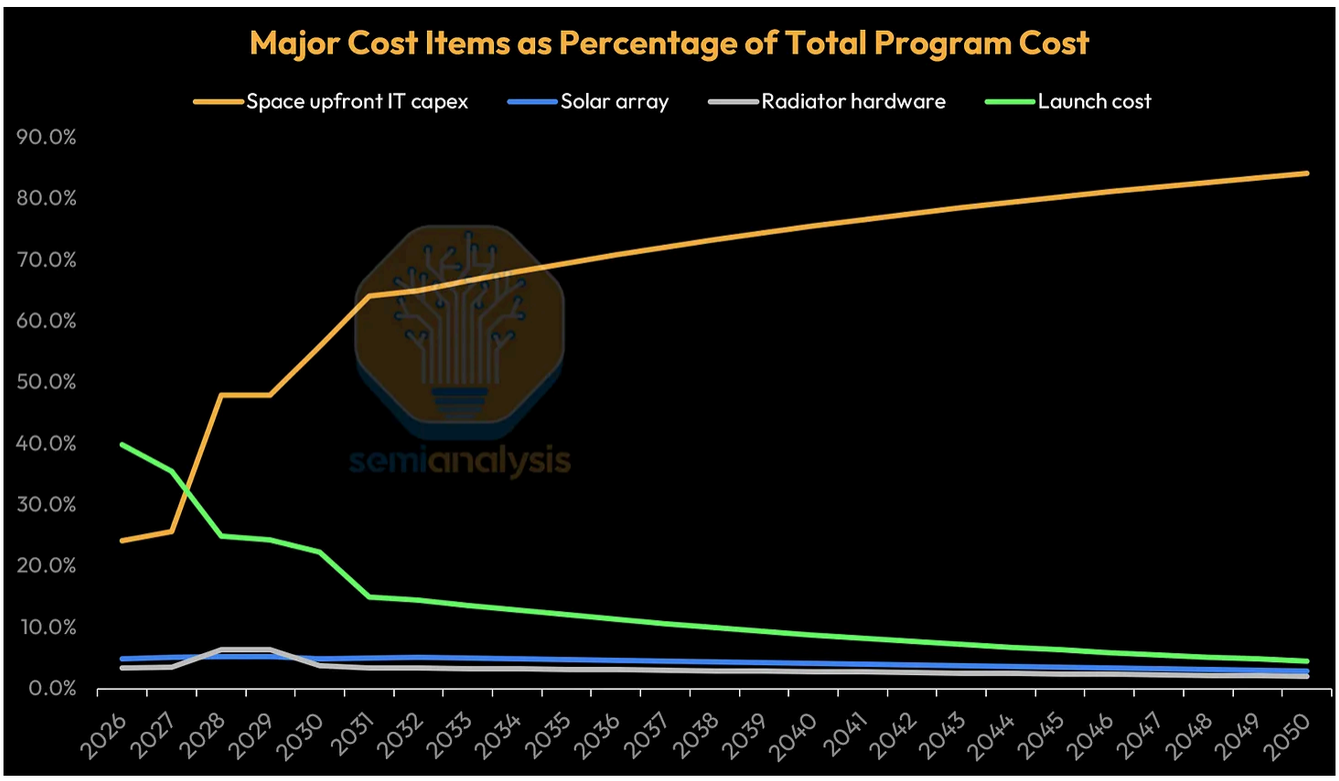
Assembly, Integration and Test Cost Buildup			
	Unit	2026	2032
Subtotal cost, excluding GPUs	$	1,254,466	17,814,433
Assembly, integration and test rate	%	15%	10%
Assembly, integration and test cost	**$**	**188,170**	**1,781,443**

Source: SemiAnalysis AI Space Datacenter TCO Model

Launch Cost Economics and Scaling: Falcon 9 vs. Starship

For the earliest years of the orbital compute program, launch cost is the largest contributor to the overall total cost of ownership for a space datacenter. At current levels (~$1,750/kg), launch costs take up 40% of total program cost for the modeled 30kW space datacenter deployed in 2026.

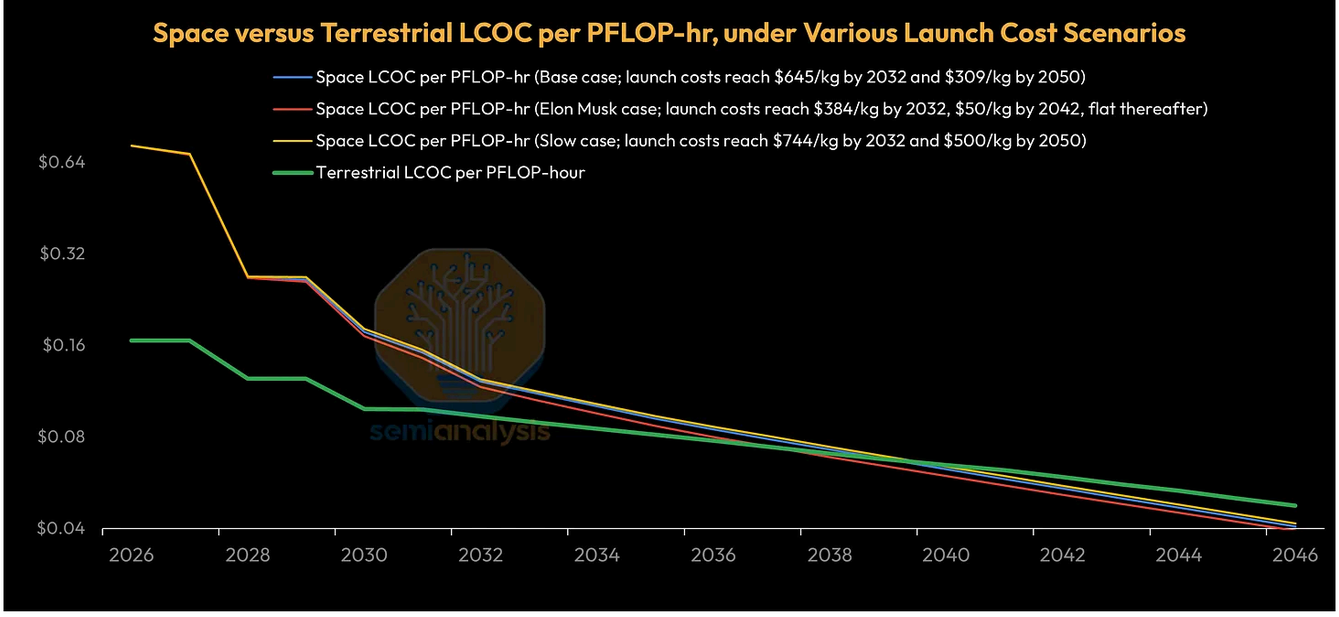
However, with the deployment of Starship in the late 2020s, we assume launch costs drop to ~$650 by 2032, meaning that launch costs drop significantly to 14.5% of total program capital costs for the ~1.5MW space datacenter deployed in 2032.



Source: SemiAnalysis AI Space Datacenter TCO Model

We consider a range of cases to display this sensitivity in LCOC per PFLOP-hr to launch costs. The chart below shows Space LCOC per PFLOP-hr in three scenarios: (1) our base case, whereby launch costs reach $645/kg by 2032 and $309/kg by 2050, (2) the aforementioned Elon Musk case where launch costs reach $384/kg by 2032, ~$80 by 2039 and $50/kg by 2042, and remain flat thereafter till 2050, and (3) a slow case whereby launch costs reach $744/kg by 2032 and $500/kg by 2050.

The irony is that, when read together with the chart above, we can see that improving launch costs has, in aggregate, very limited effects on the LCOC per PFLOP-hr results - the chart below is displayed on a log y-axis to accentuate the (relatively minor) impacts of launch costs. For the early years of 2026 to 2027 where launch costs are highest as a percentage of total program capital costs, only very small improvements can be made to the existing Falcon 9 program. By 2028 onwards, where Starship is modeled to kick in, IT capex starts to outrun launch costs relative to the total program costs, meaning that our Elon Musk aggressive case only outperforms our slow case by $0.007/hr/GPU on an LCOC per PFLOP-hr basis in 2032.



Source: SemiAnalysis AI Space Datacenter TCO Model

Nevertheless, SpaceX's innovations in this space remain necessary and ingenious. SpaceX has launched ~7,400 metric tons of cumulative mass to orbit through March 2026, accounting for >80% of global mass to orbit annually since 2023.

Falcon 9 completed 165 launches in 2025 alone, with cumulative >540 flight-proven booster launches and ~530 successful landings as of March 31st, 2026. Boosters are engineered for up to 40 flights, with a single booster demonstrating a record 34 flights (per the S-1). SpaceX uses a more conservative maximum accounting useful life of 25 flights. But each flight still requires significant ground processing. You need to crane the booster into a horizontal position, transport it to the hangar, inspect the stage, conduct vertical integration with a new second stage before moving the newly reassembled launch vehicle to the pad, fuel it, and then launch it. The total turnaround is measured in weeks.

Most of the cost reduction comes from stage reuse, and Falcon 9 reuses only one stage. The first-stage booster lands and is able to be reused, typically 20–30+ times per booster as of recent launch cycles. The second stage is expendable, so every Falcon 9 launch requires a new second stage.

The cost breakdown of a Falcon 9, based on Elon Musk's public statements (2020–2023) is as follows:

Falcon 9 Cost Breakdown

Component	Estimated Cost	% of Total
First stage / booster	~$30M	60%
Second stage / upper stage	~$9M	18%
Fairing	~$6M	12%
Integration + other	~$5M	10%
Total cost	**~$50M**	

Source: SemiAnalysis AI Space Datacenter TCO Model

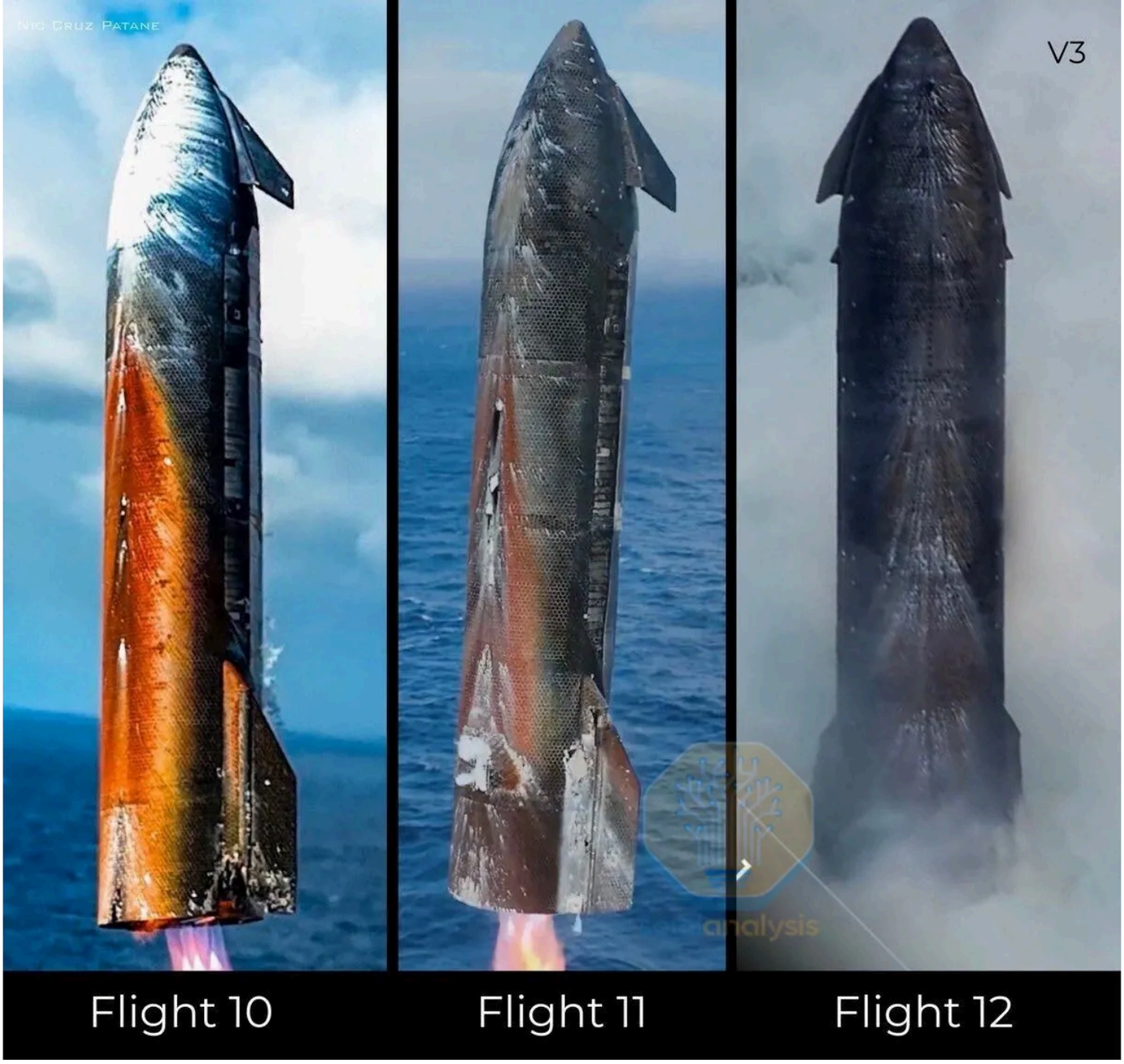
With the booster already built and paid for, the marginal cost of the first stage in a reused launch is roughly $1.2M. The expendable second stage is highest at ~$9M, followed by booster refurbishment, fairing recovery and refurbishment, propellant, and ground operations.

SpaceX charges external customers ~$67–70M per launch. Internal Starlink missions (~122 of 165 launches in 2025) cost ~$15M on the margin. Marginal cost understates the real picture, however; it excludes amortized booster build, factory overhead, engineering staff, R&D allocation, and launch infrastructure capex. SpaceX's fully loaded internal cost is roughly $31M per launch, or as low as ~$1,350–1,400/kg to LEO (we caveat again that this estimate is for fully loaded payloads, however subscale payloads are not uncommon and we assume some degree of inefficiency in our estimate of a $1,750 launch cost today). This is consistent with FY25 Space Capex of $3.8B ÷ 122 internal launches = $31M/launch.

Even with booster reuse, Falcon 9 consumes 20–30% of its hardware value on every flight, and the expendable second stage is an upper limit on how cost-effective launches may become.

In recent news, SpaceX launched Starship Flight 12 on May 22, 2026, the maiden flight of the redesigned V3 vehicle, from the new Pad 2 at Starbase. The attempt slipped a day from May 21 after a hydraulic pin issue scrubbed the countdown in the final minutes.

The mission validated the V3 architecture under engine-out conditions. The upper stage reached its planned suborbital trajectory and splashed down on target in the Indian Ocean after losing one vacuum Raptor in flight, with the remaining engines compensating. It also deployed 22 Starlink simulators through an upgraded "Pez dispenser" mechanism, the same deployment path the operational vehicle will use. The booster did not survive. Only one engine ignited for the Super Heavy landing burn, and the booster impacted the Gulf of Mexico at ~1,450 km/h.



Source: SpaceX

The headline number for payload economics is the new ceiling. V3 targets >100 tons to LEO fully reusable, with up to ~200 tons in expendable configurations. That capacity projection is what underpins the intermediate cost target referenced earlier. The upper stage's tolerance to an in-flight engine-out is a meaningful early data point for the reliability needed to fly compute payloads.

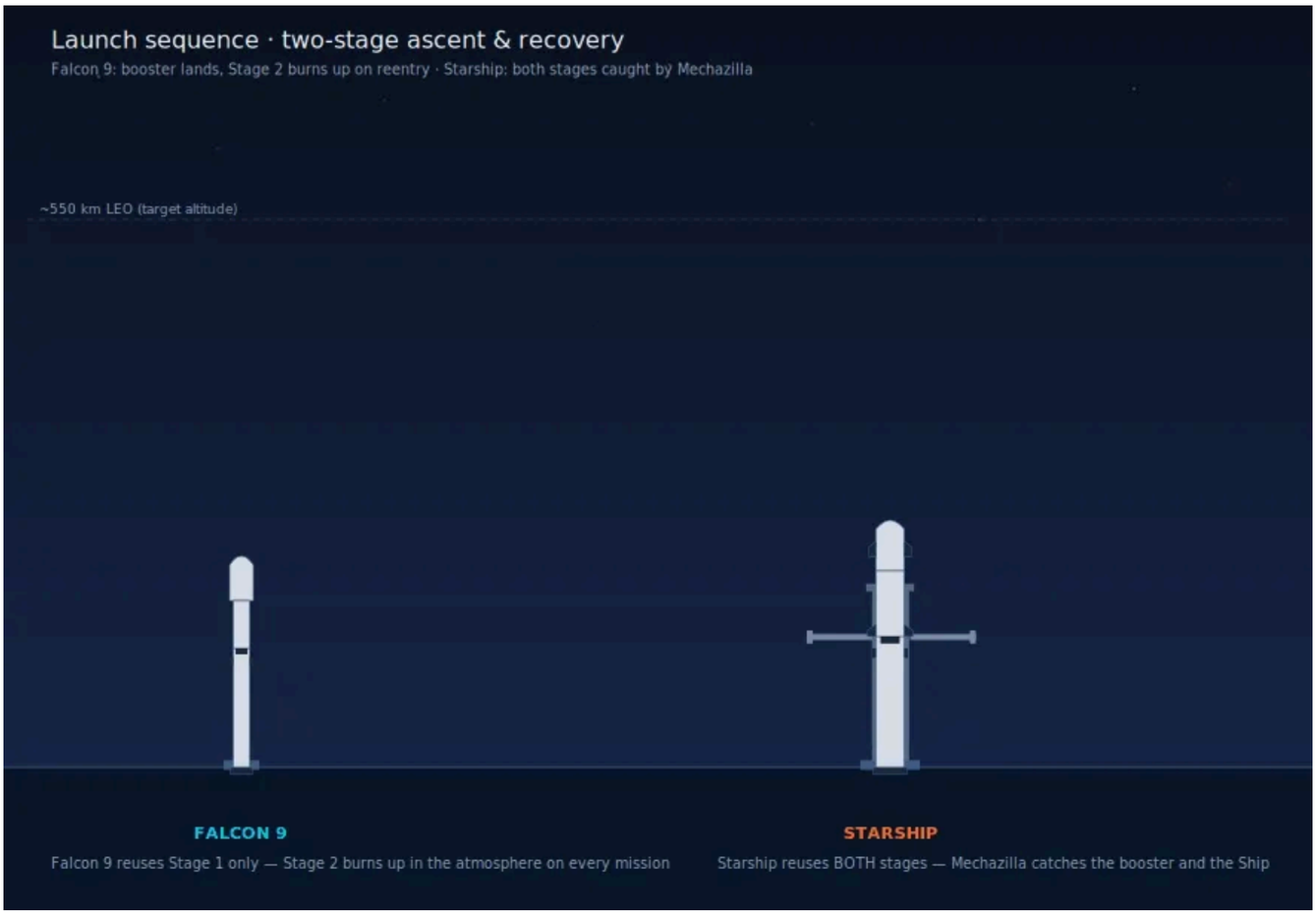
Eliminating the Expendable Floor

Starship is designed to recover everything, and do so rapidly. The Super Heavy booster returns to the launch tower and is caught by mechanical arms ("Mechazilla"), as demonstrated successfully in October 2024. The Starship upper stage is designed for the same tower-catch recovery; controlled ocean splashdowns have been demonstrated, with the upper-stage landing "within three meters of its intended landing point", and tower catches planned for 2026.

Starship's tower-catch architecture is designed for a much faster turnaround. The booster never goes horizontal. It's caught by the tower, inspected in place, restacked with a Starship upper stage, and launched again. SpaceX's stated goal is same-day turnaround, though that is unlikely to be achieved in the near-future. Even a weekly

turnaround per pad would enable 50+ flights per year per pad, and SpaceX is building multiple pads.

When both stages are recovered and reflown, the per-flight cost structure changes fundamentally.



Source: Astrocompute, SemiAnalysis

Falcon 9 versus Starship Cost Structure			
	Unit	Falcon 9	Starship
Capital cost			
First stage / booster	$M	30.0	70.0
Second stage / upper stage	$M	9.0	20.0
Fairing	$M	6.0	-
Integration + other	$M	5.0	-
Total new-build vehicle	**$M**	**50.0**	**90.0**
Reusability			
First stage / booster	# of uses	25.0	5.0
Second stage / upper stage	# of uses	1.0	5.0
Fairing	# of uses	10.0	-
Integration + other	# of uses	1.0	1.0
Effective capital cost per flight			
First stage / booster	$M	1.2	14.0
Second stage / upper stage	$M	9.0	4.0
Fairing	$M	0.6	-
Integration + other	$M	5.0	-
Total new-build vehicle	**$M**	**15.8**	**18.0**
Operating cost per flight			
Propellant + helium	$M	0.3	1.0
Booster refurb (engines + structure)	$M	1.0	2.8
Recovery ops (droneship, port, transport)	$M	2.0	2.0
Ground ops	$M	6.3	10.0
Range fees	$M	1.0	0.5
Total new-build vehicle	**$M**	**10.6**	**16.3**
Effective cost per flight	$M	26.4	34.3
Realistic delivered payload[1,2]	kg	16,000	50,000
Cost per kg to LEO	**$/kg**	**1,652**	**685**

1. Falcon 9 payload 16,000kg to 22,800kg, Starship payload up to 150,000kg.

2. For Starship early-stage reuse, we assume 5 reuses for the first and second stage, at 50,000kg payload.

Source: SemiAnalysis AI Space Datacenter TCO Model

Payload Scale - A Few Practical Examples and Future Scenarios

As mentioned, Falcon 9 delivers 16,000kg-22,800 kg to LEO in reusable configuration. Starship delivers 100,000–150,000 kg to LEO in reusable configuration (current Block 2 vehicles closer to ~40,000kg, V3 intending to close this gap), with future stretched variants potentially reaching 200,000–250,000 kg. Even at identical cost per launch, Starship's 5-7× higher payload means the amortized cost per kilogram drops proportionally.

SpaceX has built Starbase in Boca Chica, Texas as a vertically integrated factory for Starship production. The company has discussed eventual production rates of dozens of vehicles per year. At 100+ launches per year, the fixed costs of range operations,

launch infrastructure, mission control, and program management amortize more favorably than at lower launch volumes.

Musk's goal is to replicate the airline economics model, except for space. A Boeing 737 for Southwest costs ~$100M but flies 6–8 flights per day, ~350 days per year. The fixed cost per seat-mile approaches zero; the marginal cost is fuel, crew, and landing fees. Starship's aspiration is the same: a vehicle that costs $50-90M to build, flies dozens or hundreds of times, and where the marginal cost per flight is dominated by propellant.

Combined with the lower per-launch cost from full reusability, a mature Starship platform could break sub-$500/kg per launch:

Falcon 9 versus Starship Cost Per Launch

Scenario	Cost per Launch	Payload per Launch to LEO	Launch Cost per kg
Falcon 9 (external price)	~$74M	16,000 kg	~$4,625/kg
Falcon 9 (internal cost)	~$26M	16,000 kg	~$1,652/kg
Starship (early reuse[1])	~$34M	50,000 kg	~$685/kg
Starship (maturing reuse[2])	~$25M	100,000 kg	~$253/kg
Starship (mature reuse[3])	~$20M	150,000 kg	~$132/kg

1. Assumes first and stages reused 5x.
2. Assumes first and stages reused 10x, higher payload capacity.
3. Assumes first and stages reused 25x, higher payload capacity.

Source: SemiAnalysis AI Space Datacenter TCO Model

Space Opex Components

Capital costs still dominate total cost of ownership (TCO) for terrestrial datacenters at typically 75–80%, and run higher still in space given the larger upfront capital outlay, though opex still matters. For a terrestrial self-build, electricity is by far the biggest opex line; space datacenters have minimal opex because on-board solar panels supply the electricity. It is effectively an opex-for-capex swap. We will wrap up this deep dive with a quick overview of space operating costs.

Power Costs

Electricity for a ground datacenter costs on average 8.7 cents/kWh which results in around ~$25k/year for a 30 kW cluster with a 1.35 PUE and ~80% utilization. This energy cost scales linearly and compounds.

The orbital satellite's cost of energy is CapEx, captured in the solar array. The sole energy-related cost is the 2.5%/year degradation of solar cell efficiency, which is accounted for by oversizing the solar array at the beginning of life, effectively locking-in a known energy cost at deployment.

A space datacenter's power opex advantage will become more significant if power costs increase, and this is a main argument in favor of space datacenters in a future power constrained scenario.

Maintenance and Servicing

Satellite servicing is not done at scale, and is not economically viable. A servicing mission would cost more than a replacement satellite at Falcon pricing. However, **with compute hardware as the highest cost bucket over the very long-run**, you must have robust on-orbit robotics servicing and maintenance capabilities.

On the ground, servicing is a lot simpler. You hot-swap failed drives, replace blown PDUs, reseat transceivers that are flapping and RMA broken GPUs. For space however, failed hardware will be stuck in place. The platform must therefore be designed with redundancy sufficient to survive the full 5-year mission. Our model uses a 5% annual COTS (commercial off the shelf) failure rate (conservative as Starlink achieves 2-3%), which is addressed by over-provisioning rather than servicing.

Bandwidth and Spectrum

Terrestrial AI-cluster bandwidth is primarily an economic scaling problem. A DGX B300-class 8-GPU node can expose up to 8 × 800 Gb/s InfiniBand/Ethernet interfaces depending on configuration, so the expensive network is usually the local GPU/cluster fabric rather than WAN egress.

In space, the hard constraint is different. The constellation still needs local networking and ISLs, but space-to-ground egress is physically and operationally constrained by pointing, spectrum, ground-station availability, weather, atmospheric loss, and licensing. A 20 Gbps Ka-band downlink with 17.1 Gbps effective average capacity gives ~185 TB/day per satellite, or ~185 PB/day across a 1,000-sat constellation. A typical inference workload (e.g. 500 GPUs/sat × 5K tok/s × ~100 bytes/tok $\approx$ 22 TB/day per sat, or ~22 PB/day fleet-wide) sits at roughly 12% utilization of that budget — meaning Ka-band is sufficient for inference with ~8× headroom, no optical ground network required. Optical ground links would expand capacity 10–100× and become necessary for training workloads (large dataset ingest, model-checkpoint egress) or multimodal inference (video, high-resolution images, very large context windows), but they add significant ground-segment cost and weather-dependent availability. The

inference-only case is correspondingly cheaper on the ground side, which is one of the structural reasons inference is the natural first orbital workload.

Spectrum and regulatory cost are strategically important but not a dominant dollar cost. The main RF burden is the Ka-band space-to-ground link; optical ISLs operate in unregulated wavelengths and carry far less spectrum burden, though they still sit inside broader satellite licensing, safety, export-control, and operational review. Raw filing fees are modest: ITU cost-recovery filings run roughly ~$30K per satellite network filing, FCC NGSO application fees run ~$18K per filing, and the FY2025 FCC annual regulatory fees list large NGSO constellations at ~$1.92M per authorized system. The larger recurring cost is legal, RF-engineering, coordination, market-access, and compliance staff. A reasonable model assumption is ~$10–30M upfront and ~$5–15M/year recurring for spectrum/regulatory execution. Amortized across a 1,000-sat fleet, this works out to roughly $5K–$15K per satellite per year, or well under 0.1% of per-sat capex — minor compared with satellite manufacturing, launch, compute hardware, power, thermal, and replenishment.

Ground Operations

The ground segment for an inference constellation is a global network of Ka-band gateway stations sized for downlink with weather and visibility diversity, plus the standard mission-operations stack (24/7 anomaly response, flight dynamics and conjunction screening, OTA software management). The downlink uses Ka-band in the 17.8–20.2 GHz commercial range with dual-polarization phased-array antennas on the satellite (the same architecture SpaceX uses for Starlink V2 gateway links) paired with ~3m tracking dishes on the ground, delivering ~20 Gbps peak per sat with adaptive coding to ride out atmospheric attenuation. Each LEO pass over a given gateway is only ~5–10 minutes long, so the constellation relies on optical inter-satellite links to store-and-forward traffic to whichever satellite is currently over a ground station. This is what brings the per-sat effective average down from 20 Gbps peak to ~17 Gbps sustained. Ka-band is also rain-fade-sensitive, which is why the network needs geographically distributed sites for 99%+ aggregate availability. Upfront ground-segment capex runs around $45M at fleet launch and scales toward ~$160M as bandwidth demand grows with the constellation. Ongoing operations (staffed NOC, remote hands at each gateway, fleet engineering, customer support, and SLA management) add some tens of thousands of dollars per satellite per year, scaling with platform complexity.

The structural advantage on the ground side is that the work is mostly fixed-overhead: a single mission ops center and a handful of regional gateway sites can support a fleet of hundreds to thousands of sats, so the per-satellite operating cost amortizes down quickly as the constellation grows. Unlike a terrestrial datacenter, there are no on-site technician swaps to schedule, no spare-parts inventory at each rack, and no facility-level maintenance. Once a satellite is on orbit, the ground-side touch is mostly software, telemetry, and licensing rather than physical labor. The result is that ground

stations and operations together run roughly under $100M a year for a 1,000-sat constellation at the 2026 baseline, scaling with platform complexity as per-sat cluster cost grows. It's meaningful, but a small fraction of total TCO.

Launching our AI Space Datacenter TCO Model

Today, we are also launching our AI Space Datacenter TCO Model into General Availability. The model provides a first-principles, system-level framework for evaluating orbital compute economics, engineering constraints, and supply-demand dynamics across both terrestrial and space-based infrastructure.

It spans from launch vehicle physics and thermal rejection limits to AI demand curves and GPU-level cost of ownership, enabling users to stress-test when and how compute may transition off Earth. Our model spans 2026 - 2050, with dynamic scenario modeling driven by user-controlled assumptions.

The model will answer the question of if and when space datacenters will become economical and will show the scenarios with respect to end demand and costs on Earth that make Space a viable deployment option.

Please reach out to sales@semianalysis.com to find out more!

Previous

A guest post by

Pranav Myana

hello

Discussion about this post

LIKE REPLY SHARE

Dan 5h

SEU and SEFI problems are SUBSTANTIALLY more difficult to deal with for space data centers than a communications satellite. The author must've never talked to anyone who's done board-design for space. This entire analysis is trash.

LIKE REPLY SHARE

1 more comment...