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

209 10 19 Share

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

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.

Users of our AI Space Datacenter TCO Model can see a first-principles, system-level framework for evaluating orbital compute economics, engineering constraints, and supply-demand dynamics across both terrestrial and space-based infrastructure.

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

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

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

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!

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

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.

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

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

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.

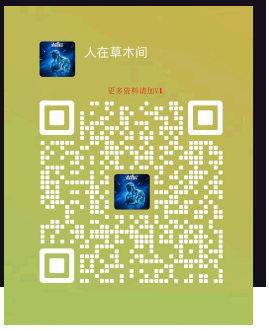
Source: SemiAnalysis AI Space Datacenter TCO Model

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

Source: SemiAnalysis AI Space Datacenter TCO Model

Source: SemiAnalysis AI Space Datacenter TCO Model



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

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

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.

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?

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 chip deployment, whether deployed on Earth or in Space.

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,

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

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.

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.

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.

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.

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 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

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**.

Source: SemiAnalysis AI Space Datacenter TCO Model

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.

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.

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.

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

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

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.

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.

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.

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

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.

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

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.

Source: SemiAnalysis AI Space Datacenter TCO Model

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.

This post is for paid subscribers

Subscribe

Already a paid subscriber? Sign in

© 2026 Dylan Patel · Privacy · Terms · Collection notice

Start your Substack

Get the app

Substack is the home for great culture