Where Should AI Compute Be Built in 2029–2030?

Geographic, Energy and Industrial Frontiers of Terrestrial, Stratospheric and Orbital Data Centers

A quantitative prospective study. Evidence cutoff: September 25, 2026. Decision horizon: commissioning in 2029 and 2030. Reproducible edition with data, assumptions, equations and figures.

Abstract

The optimal location for artificial intelligence compute cannot be inferred from a country's average electricity cost, outdoor temperature or available solar irradiance alone. It emerges from the intersection of five capabilities: supplying firm power, rejecting heat, transporting data, replacing equipment and delivering the facility before the customer needs it. This study compares grid-connected terrestrial architectures, dedicated electricity supply, coastal facilities, stratospheric aerostatic platforms at altitudes of 20–25 km, and low-Earth-orbit compute. It explicitly examines the SpaceX industrial proposal and Google's Suncatcher research. The initial hypothesis of an inherent stratospheric advantage is replaced by an investigation of the conditions under which each architecture is competitive.

The IEA central scenario raises global data center electricity consumption from 485 TWh in 2025 to 950 TWh in 2030. This growth does not establish a global energy shortfall: the decisive constraints may be local, hourly and schedule-dependent. Fifteen regions are examined using primary documents, while future prices and lead times are explored over explicitly hypothetical ranges. No simulation range is presented as a confidence interval. The techno-economic model uses a common service unit of one nominal MW of IT power, an initial active-load compute efficiency of 5×10^{11} FLOP/J, 75% utilization and six operating years. It selects between hardware replacement after three years and retention for six years.

In the hypothetical central case, levelized costs are 0.6301, 0.6964 and 0.8976 USD per 10^{18} useful FLOP for grid-connected terrestrial, low-Earth-orbit and stratospheric compute, respectively. These values are neither cloud tariffs nor price forecasts; they put the assumptions on a comparable basis. Orbital compute becomes competitive against the reference terrestrial site when its electricity price exceeds approximately 177 USD/MWh, assuming launch at 200 USD/kg and a specific mass of 16.97 kg/kW IT. This favorable result remains conditional on the assumed availability, networking, qualification and manufacturing performance, together with a nearly continuously illuminated trajectory that requires no battery for long eclipses. At 20 kg/kW and 100 t per mission, deploying one orbital GW between 2028 and 2030 would require approximately 136 launches annually if half of all flights were allocated to compute and 98% of payloads were delivered.

For the stratosphere, an optimistic coupled physical model at 22 km yields, per continuous MW IT, a 371 m aerostat, 165 t excluding helium and 18.6 MWh of nominal battery capacity. The initial helium inventory for a one-GW fleet would reach approximately 26.5 kt, compared with estimated annual global production of about 32.2 kt in 2025. Hydrogen, intermittent workloads and alternative architectures may alter this constraint, but they define different systems requiring validation. The scientific conclusion is conditional: several locations may become optimal for different applications, and some orbital designs may achieve an economic advantage before they can serve a substantial share of global demand. The investment thesis should target a portfolio of deliverable capacity increments rather than declare a universally optimal location.

Scientific Status and Reading Guide

This document is a theoretical, prospective study based on public sources and first-order models. It does not claim to result from two years of testing, a university affiliation or industrial qualification. Its level of detail is intended to provide a verifiable scientific working basis; missing experimental observations are identified as such. Quoted industrial figures retain their status as announcements when they do not describe demonstrated performance.

Four categories are distinguished: **documented fact**, **industrial target**, **exploratory assumption** and **calculated result**. The CSV files preserve these distinctions. An industrial target may inform a favorable scenario without becoming an established fact. Financial parameters are expressed in constant 2026 USD, excluding general inflation, and the model discount rate is interpreted as real. Future commercial offers must be converted to this basis, or the entire model must be converted to nominal currency. No future exchange rate is predicted.

The main report presents the methodology, results and limitations. Figures are supplied in PNG and SVG formats, together with the data and programs. An appendix contains a standalone program reproducing the three requested benchmarks: thermal performance, deployment and cost–latency. The remaining figures can be regenerated using the scripts in the `code` directory. The source ledger distinguishes historical measurements from prospective trajectories.

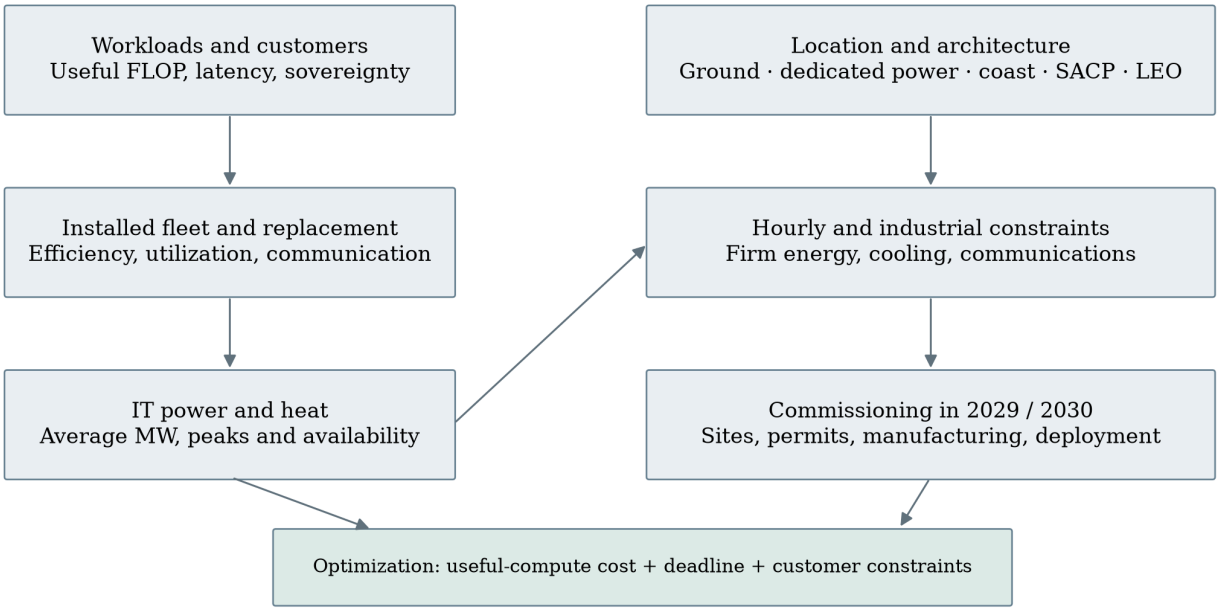
Research Questions

The first question concerns the amount of useful compute actually required in 2029–2030 and how improvements in efficiency change the associated power demand. The second concerns the locations where that power can be contracted and brought into service. The third identifies the cost, mass, service-life and latency thresholds that open a competitive domain for orbital or stratospheric compute. The fourth examines whether these domains can scale from a demonstrator to 100 MW and then to a GW. A solution that is economically attractive at a small scale may lose its advantage at the next scale because the scarce resource changes.

The desired result is therefore not an opinion on “Earth or space.” It is a decision function:

$$(i^*, a^*, s^*) = \arg\min_{i,a,s} C_{useful}(i, a, s \mid w, t, \Theta),$$

subject to service, deadline and capacity constraints. Here, i denotes location, a architecture, s system sizing, w workload, t date and Θ the set of assumptions. Changing the workload, date or replacement cost may change the optimal location. Explicitly identifying these transitions is more informative than a single ranking.



F01. Causal structure of the model. A location is admissible only after physical balances, scheduling and customer constraints are satisfied. Methodological diagram with no observed values.

Common Framework: Comparing the Same Service over the Same Time Horizon

Functional Unit and System Boundaries

Useful compute is defined at a common numerical precision, output quality and workload. A dense tensor FLOP, an operation exploiting advertised sparsity and a generated token are not interchangeable units. Converting FLOP into revenue requires an application-level model. We use a common reference efficiency to isolate the value of location; it does not predict the specifications of any particular accelerator in 2030.

For nominal IT power P, active-load efficiency η, utilization u, availability a and relative application efficiency χ, throughput is

$$\dot{F}_{useful} = P\eta u a \chi.$$

The central scenario combines one MW IT, an active-load efficiency of 500 billion FLOP/J, 75% utilization, 99.5% availability and a relative efficiency of one. It yields 3.73125×10^{17} FLOP/s. A block of 10^{18} FLOP therefore corresponds to approximately 2.68 seconds of service from that MW. The monetary unit USD per 10^{18} FLOP describes an amount of computation, not an exaflop per second. Values in the report must be multiplied by 10^{-18} to obtain USD/FLOP.

Average electrical power is not uP when machines consume power at idle. With $f_0 = 0.20$:

$$\overline{P}_{IT} = P[u + (1-u)f_0] = 0.80P.$$

The reference facility therefore consumes an average of 0.92 MW per nominal MW IT at a PUE of 1.15. Airborne and orbital platforms are sized to support full load, including nighttime operation where continuity is required. This reserves electrical capacity that may temporarily be underutilized. That reserve cannot simultaneously be counted as a service guarantee and omitted to claim a mass saving. A strictly opportunistic compute architecture is a different product and requires its own hourly profile.

Central-Case Parameters

Parameter	Central value	Status and interpretation
Target commissioning year	2029 or 2030	Decision horizon, not a guarantee
Discounted operating horizon	6 years	After commissioning
Real discount rate	10%/year	Common assumption; sensitivity analysis required
IT cost	18 M USD/MW IT	Assumed 2030 value, not a supplier quote
Initial active-load efficiency	5×10^{11} FLOP/J	Common service convention
Utilization / idle fraction	0.75 / 0.20	Distinct assumptions
New-generation efficiency improvement	15%/year	Realized only when equipment is replaced
Replacement options	3 or 6 years	Discrete economic choice
Terrestrial infrastructure	8 M USD/MW IT	Excludes IT; assumed value
Terrestrial electricity	80 USD/MWh	Assumed all-in delivered energy price
Terrestrial PUE	1.15	Assumed weighted annual value
Fixed terrestrial OPEX	0.8 M USD/MW/year	Assumption, excluding energy
Terrestrial / SACP / LEO availability	0.995 / 0.96 / 0.97	Assumptions, not demonstrated SLAs
Fixed SACP / LEO OPEX	2.0 / 1.1 M USD/MW/year	Operating assumptions
Orbital launch	200 USD/kg	Scenario value, not a contractual price
LEO radiating surface	350 K; $\varepsilon = 0.90$	Design assumptions
LEO absorbed external flux	100 W/m²	Thermal parameter requiring qualification
Orbital illumination fraction	0.98	Favorable architecture; eclipses must be checked
Stratosphere	22 km; 12 h night	Standard atmosphere and solar scenario
SACP battery	250 Wh/kg at system level	System assumption, not cell-level performance

The intended symmetry is not identical line items. Terrestrial compute incurs an electricity bill; orbital compute requires panels, radiators and launch services. Comparability rests on complete system boundaries and an identical service. A common 10% discount rate isolates technical differences. Actual financing may require a higher return for an architecture with no operating track record; a risk premium should be added in sensitivity analysis without presenting a subjective judgment as market data.

Constraints and Formal Optimization

Let x_{iawt} denote the power allocated to workload w at site i, using architecture a, at time t. Costs are meaningful only within the feasible set $\mathcal{F}$:

$$\sum_{a,\,w} b_{ia} x_{iawt} \le K_{it}^{firm}, \qquad L_{iaw} \le L_{w}^{max}, \qquad x_{iawt} \le M z_{iat}.$$

The coefficient b converts IT power into demand at the supply point; K^{firm} is demonstrably available capacity, and z indicates industrial readiness. For an orbital fleet, the grid-connection constraint is replaced by constraints on onboard generation, satellite production and delivered mass. For SACP, buoyancy, storage, weather and operating authorization add further constraints.

Service demand in scenario ω requires

$$\sum_{i,\,a} x_{iawt}\eta_{iat}u_{iawt}a_{iat}\chi_{iaw} + s_{wt} \geq D_{wt}^{\omega},$$

where s represents unserved demand. Its economic penalty depends on the customer. Annual budgets, network capacities, water and carbon constraints introduce further inequalities. The objective may be discounted total cost, margin or maximum regret. These objectives yield different answers; the commercial decision must select the one that matches the contracts.

This study does not claim to solve a global allocation problem with unknown inputs. It solves reproducible physical and economic subproblems and then maps candidate regions for investigation. Without grid-connection offers, hourly profiles and customer contracts, claiming a global optimum would be unjustified. The work derives its value from measurable thresholds and experiments that can eliminate branches of the decision problem.

Demand, efficiency, and power actually available in 2029–2030

Growth in digital computation does not demonstrate an intractable global electricity shortage. It creates competition for a more specific resource: power that can be delivered continuously, secured contractually, and connected on time. This distinction determines the relevance of future architectures. Moving a data center into the stratosphere or orbit can eliminate its connection to the terrestrial grid; it nevertheless transfers its energy requirements to onboard generation, storage, and distribution infrastructure. The relevant comparison therefore encompasses the complete system required to deliver an equivalent computing service.

The 2029–2030 horizon requires a distinction between three quantities: annual energy consumption, power required during critical hours, and industrial capacity that can actually be commissioned. A demand forecast describes the first or second. A campus announcement often describes only an ambition concerning the third. Adding them together creates an appearance of precision without a consistent physical balance. The data and scenarios presented here are current as of 25 September 2026; each result retains the publication date of its source.

1. Define the load before forecasting it

A facility's electrical energy consumption is the integral of its power draw:

$$E_{\mathrm{fac},\,y} = \int_{\mathcal{T}_y} P_{\mathrm{fac}}(t)\,dt, \qquad \overline{P}_{\mathrm{fac},\,y} = \frac{E_{\mathrm{fac},\,y}}{T_y}.$$

For a year of 8,760 hours,

$$1\,\mathrm{GW_{average}} = 8.76\,\mathrm{TWh/year}.$$

Average power is neither peak demand nor requested connection capacity. If K_{conn} denotes the latter, its electrical load factor is

$$\lambda_{\mathrm{conn}} = \frac{E_{\mathrm{fac}}}{K_{\mathrm{conn}}T}, \qquad K_{\mathrm{conn}} = \frac{E_{\mathrm{fac}}}{\lambda_{\mathrm{conn}}T}.$$

This factor combines ramp-up, design margins, redundancy, and operating profile. It does not measure GPU utilization. A lightly utilized machine can continue consuming power at idle; a heavily utilized machine can operate in a building whose halls remain largely empty.

Quantity	Unit	Meaning	Confusion to avoid
Facility energy E_{fac}	TWh/year	Annual consumption, including auxiliaries	Treating it as AI consumption alone
Average power $\overline{P}$	GW	Energy divided by duration	Treating it as firm power
Rated IT power K_{IT}	MW	Electrical capacity of computing equipment	Treating it as connection capacity
Connected capacity K_{conn}	MW	Entitlement and design rating at the delivery point	Treating it as useful throughput
Useful throughput $\dot{C}$	FLOP/s or tasks/s	Work actually delivered	Treating it as manufacturer-rated peak throughput

Defining PUE for the specified electrical system boundary,

$$P_{\text{fac}}(t) = \text{PUE}(t) P_{\text{IT}}(t).$$

The approximation $E_{\text{fac}} \simeq \text{PUE}\, E_{\text{IT}}$ requires either constant PUE or an average correctly weighted by IT energy. A forecast already expressed at the facility level must not be multiplied by this coefficient a second time. For an onboard platform, propulsion, communications, electrical conversion, and storage also require an explicit system boundary: arbitrarily placing them outside the “data center” would artificially improve its efficiency.

2. Relate deliveries, installed stock, and consumption

An industrial projection must reconstruct the stock of equipment from deliveries, installation delays, and retirements. For equipment family j, a cohort model is written as

$$N_j(t) = \int_{-\infty}^{t} D_j(\tau)\, S_j(t-\tau)\, d\tau,$$

where D_j represents commissioning and S_j the fraction of a cohort still operating at a given age. Orders are not commissioned installations. Their conversion depends simultaneously on accelerators, memory, internal networking, buildings, and energy. A discrete model preserves this distinction:

$$N_{j,y+1} = N_{j,y} + D_{j,y+1} - R_{j,y+1}.$$

In the absence of detailed measurements, an affine power approximation can provide a transparent scenario:

$$P_j(u) = P_{j,0} + (P_{j,1} - P_{j,0})u, \qquad 0 \le u \le 1.$$

This is not a universal law of accelerator behavior. It nevertheless shows why reducing utilization does not reduce energy proportionally: $P_{j,0}$ remains. Introducing availability a_j and realization efficiency ξ_j, delivered useful computation becomes

$$\dot{C}_j = N_j a_j u_j \xi_j F_{j,\max}, \qquad \eta_{\text{eff}} = \frac{\sum_j \dot{C}_j}{P_{\text{fac}}}.$$

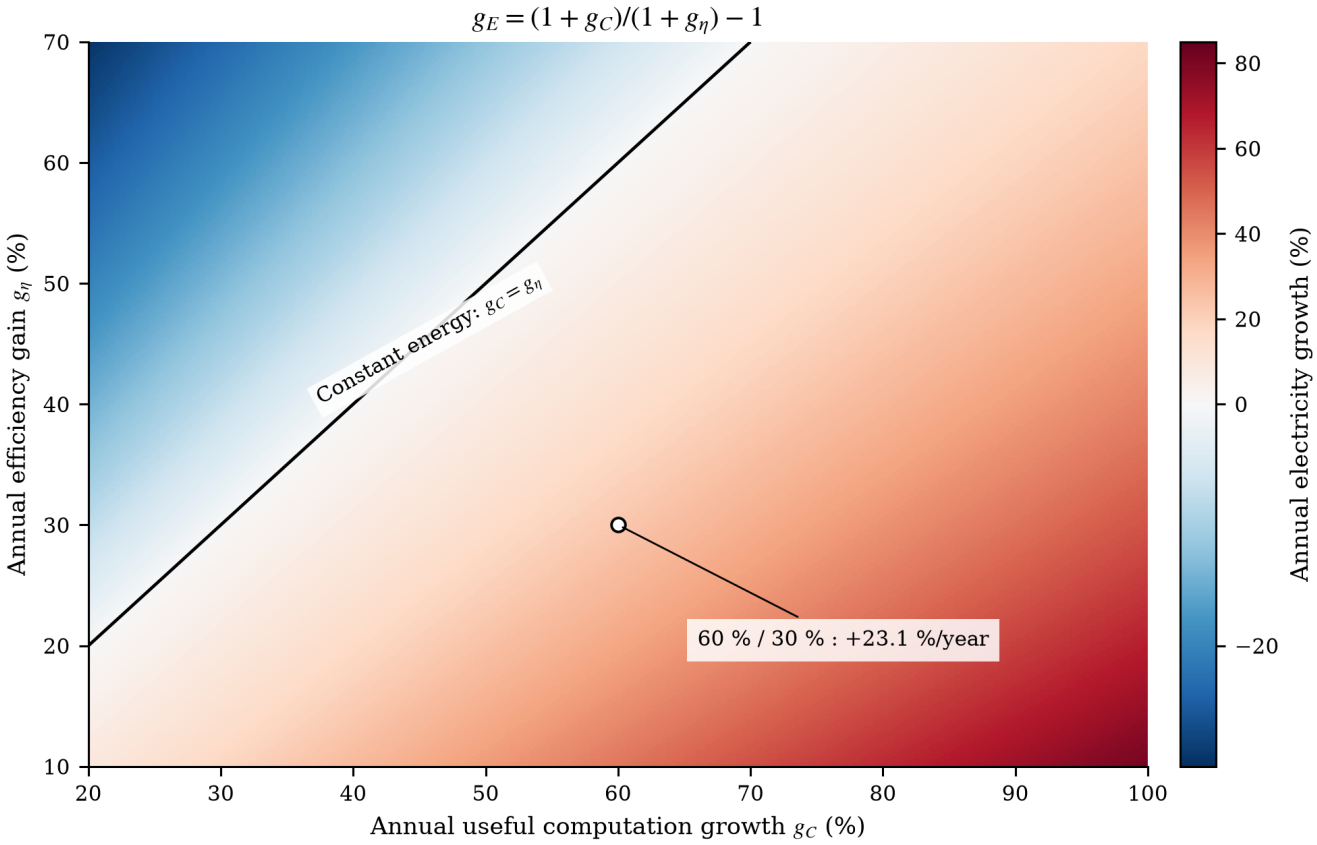
$F_{j,\max}$ must be defined using identical numerical precision, sparsity, and operation-counting conventions. Inter-accelerator communications, memory transfers, and repeated computation alter ξ_j. An

improvement in peak FLOP/s per watt therefore does not imply an equivalent improvement in distributed training or latency-constrained inference.

For annual useful work C_y and a consistently defined average efficiency η_y, the identity $E_y = C_y/\eta_y$ yields

$$\frac{E_{y+1}}{E_y} = \frac{1+g_{C,y}}{1+g_{\eta,y}}, \qquad g_{E,y} = \frac{1+g_{C,y}}{1+g_{\eta,y}} - 1.$$

Thus, an illustrative annual increase of 60% in computation and 30% in efficiency still produces $1.6/1.3 - 1 = 23.1\%$ growth in electricity consumption. Conversely, with stable demand, efficiency gains reduce energy consumption. This accounting result predicts neither future applications nor rebound effects: it identifies the variables that must be compared. Expansion into more complex tasks can absorb per-operation savings; algorithmic substitution can reduce the computation required per task without changing the desired quality.



F04. Exploratory accounting identity, not a forecast: E=C/η; therefore (1+g_E)=(1+g_C)/(1+g_η). Useful work, numerical precision, and electrical system boundary are constant; η can include PUE when consumption is measured at facility level. Rebound effects, model evolution, utilization, and the workload mix are not forecast separately. A reduction in joules per operation does not guarantee lower total consumption when computation grows faster.

3. From published anchors to 2029–2030 scenarios

The IEA projects global consumption by all data centers to increase from 485 TWh in 2025 to 950 TWh in 2030. Its AI-oriented subset grows faster; the 485 TWh therefore does not represent AI alone. IEA, April 2026.

The US LBNL report entitled *2025 Update*, actually published in June 2026, uses 192 TWh in 2024, 464 TWh in 2028, and 649 TWh in 2030 in its reference case. It covers facilities excluding cryptocurrency operations, with auxiliaries included. Its combined scenarios reach 521–843 TWh in 2030; 2024 is the model's last historical year. LBNL, pages 8–10.

These references do not form a reconciled statistical system. They rely on different vintages, industrial assumptions, and analytical treatments. Subtracting the LBNL US forecast from the IEA global forecast

would fabricate a “rest of the world” for which no institution has established the underlying assumptions. The two panels should therefore remain separate, followed by sensitivity analysis within each scope.

This study supplements the central global endpoint with two author-defined exploratory endpoints of 750 and 1,400 TWh in 2030. They test infrastructure decisions; they are neither IEA scenarios nor distribution quantiles. Between two anchors E_a and E_b, geometric interpolation is

$$E_y = E_a\left(\frac{E_b}{E_a}\right)^{\frac{y-a}{b-a}}, \qquad \bar{g} = \left(\frac{E_b}{E_a}\right)^{1/(b-a)} - 1.$$

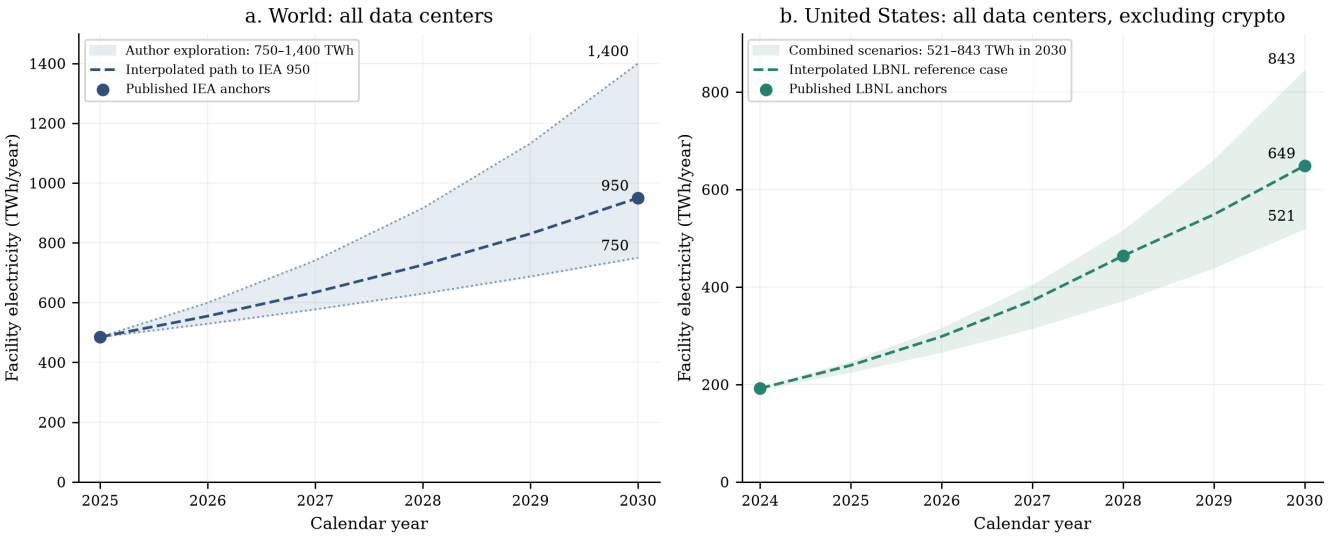
Global trajectory	Status of the 2030 value	2030, TWh/year	CAGR 2025–2030	Interpolated 2029, TWh/year
Moderate	Author-generated exploration	750	9.11%	687.38
Central	IEA anchor	950	14.39%	830.48
Accelerated	Author-generated exploration	1,400	23.62%	1,132.53

This convention facilitates reproduction; it does not establish a smooth industrial trajectory. Linear interpolation would give 857 TWh for the central global 2029 point, compared with 830.48 geometrically. Between the US points for 2028 and 2030, it would give 556.5 TWh, compared with $\sqrt{464 \times 649} = 548.76$ TWh. These differences describe only the graphical interpolation convention. Uncertainties in deployment, usage, and available power are separate and potentially larger.

A consistent trajectory then separates desired demand from achievable deployment. The former depends on applications, prices, and customer budgets; the latter depends on deliveries and infrastructure. Demand exceeding capacity does not immediately become consumed energy: it can produce task queues, higher prices, geographical relocation, or delayed projects. For service class k, one can write

$$C_{k,\,\mathrm{served}}(t) \le \min\left[C_{k,\,\mathrm{demanded}}(t), C_{k,\,\mathrm{capacity}}(t)\right].$$

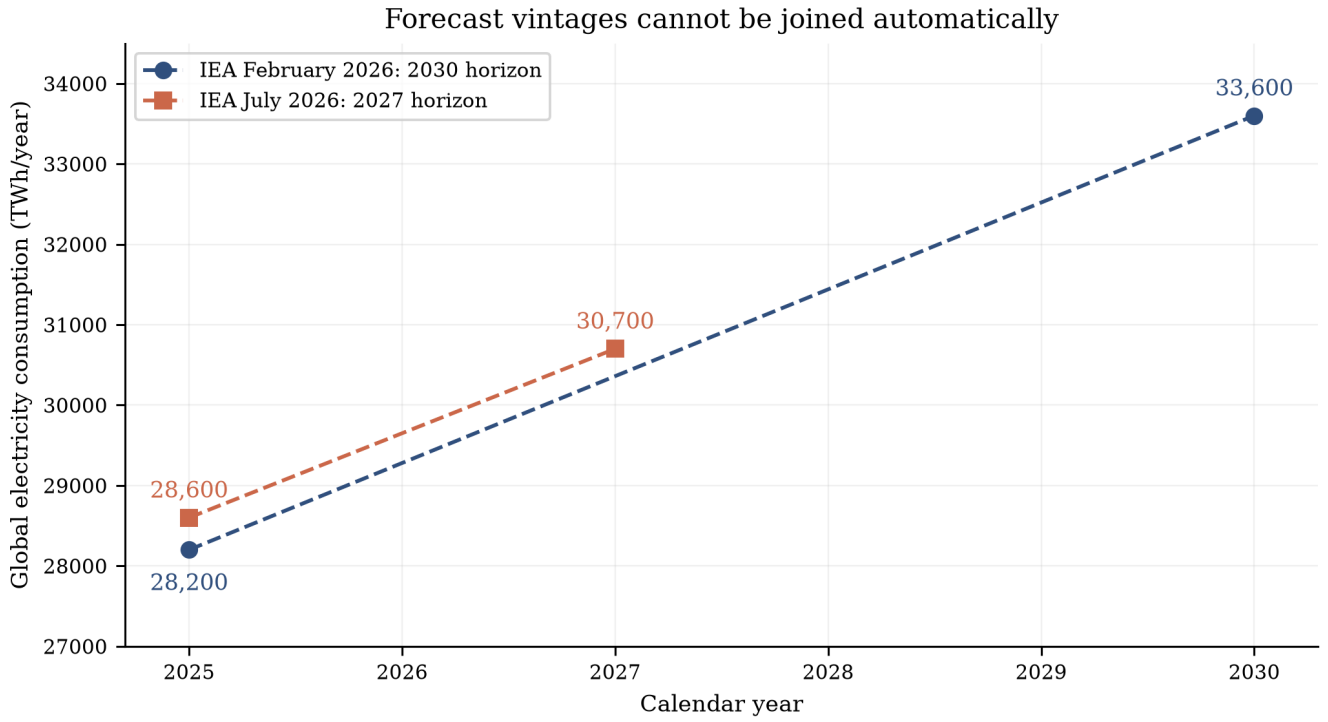
This bound requires appropriate granularity. A surplus of machines dedicated to an unsuitable numerical precision or network topology does not necessarily satisfy demand in another class. For interactive inference, the constraint may concern a latency quantile; for training, a model delivery date. An “accelerated” scenario must therefore specify its additional investments and schedule, rather than merely increase the final number of TWh. A “moderate” scenario can reflect higher efficiency, lower demand, or capacity rationing; these three mechanisms have different commercial consequences despite identical consumption.



F02. Sources: IEA Key Questions on Energy and AI, 16 April 2026, and LBNL 2025 Update, 18 June 2026. All data centers, total facility energy; excluding crypto for LBNL. World: 485 TWh (2025) and 950 (2030) are IEA anchors. The endpoints of 750 and 1,400 are author-generated explorations, without probabilities or attribution to the IEA. United States: 192 (2024), 464 (2028), 649 (2030), and endpoints of 521/843 in 2030 are published by LBNL. All annual curves between anchors are author-generated geometric interpolations. The two panels do not form a reconciled geographical balance. Sources: source ; source

4. Global energy supply expands; its location remains constraining

The February 2026 IEA forecast increases global electricity consumption across all uses from 28,200 TWh in 2025 to 33,600 TWh in 2030. The July update revises the 2025 anchor to 28,600 TWh and forecasts 30,700 TWh in 2027. An unlabeled curve using the revised baseline and the previous growth rate would be an author-generated construction improperly presented as an institutional forecast. IEA, February; IEA, July.



F03. Global consumption, all uses. Electricity 2026 (February): 28,200 TWh in 2025 and 33,600 in 2030. Mid-Year Update (23 July 2026): 28,600 in 2025 and 30,700 in 2027. Segments connect only the anchors; they do not represent intermediate observations. No July forecast for 2030 or rebased extrapolation is attributed to the IEA. Sources: source ; source

On the supply side, the February scenario forecasts approximately 1,050 TWh of average annual growth in renewable generation over 2026–2030, including more than 600 TWh from solar. The share of low-carbon generation reaches 50% in 2030, compared with 42% in 2025. IEA, Supply.

The increase in data center consumption from 485 to 950 TWh adds 465 TWh to annual consumption at the end of the period, equivalent to

$$\Delta \overline{P}_{\mathrm{DC}} = \frac{465}{8.76} = 53.08\,\mathrm{GW}.$$

This comparison refutes the idea that growth in computation alone would mechanically exhaust all possibilities for global electricity expansion. It reserves no generation for data centers: industrial electrification, mobility, buildings, and other uses grow simultaneously. Moreover, the two quantities compared are aggregates from separate publications, rather than a coordinated investment program.

The change in an annual flow must be distinguished from its cumulative value over time. If renewable generation increases by δE each year, then

$$E_{2030} - E_{2025} = 5\delta E, \qquad \sum_{k=1}^{5} (E_{2025+k} - E_{2025}) = 15\delta E.$$

The first term is a difference between annual generation volumes; the second is cumulative additional energy over five years. Confusing these terms overstates or understates supply capability. Above all, a TWh of solar electricity generated at noon in one region does not automatically supply a campus elsewhere during a windless night.

5. Regional pressures differ substantially

PJM forecasts a total summer peak of 171,530 MW in 2029 and 183,008 MW in 2030. These figures concern the entire grid, not data centers alone. Changes in its project verification methodology underscore the need to screen a load queue before incorporating it into a resource adequacy assessment. PJM, January 2026.

ERCOT illustrates the potential gap between applications and completed projects. Its preliminary April 2026 compilation reaches 367,790 MW in 2032, but the operator states that this level exceeds expected loads and does not predict actual construction. Treating it as certain firm demand would directly distort investment requirements. ERCOT.

In Australia, the August 2026 AEMO/Oxford reference case projects NEM data center consumption to increase from 5.1 TWh in FY26 to 15.5 TWh in FY30, with alternative scenarios of 10.5 and 20.1 TWh in FY30. These are fiscal years. Concentration in New South Wales and Victoria implies local constraints concealed by the national average. AEMO/Oxford.

France presents a different situation: 547.5 TWh generated and 92.3 TWh of net exports in 2025. These exports already serve foreign consumers; they are not a freely available stock. RTE also estimates data center consumption at 15–20 TWh in 2030 in its rapid decarbonization scenario, while approximately 18 GW of connections had been reserved by May 2026. The gap between energy consumption and reservations precludes a naive conversion of announcements into continuous operation. RTE, electricity review; RTE, data centers.

These observations do not support ranking countries by average spot prices. For 2029, a less conspicuous but already equipped node can outperform a cheaper territory whose reinforcement remains unfunded. Connection, fiber, permits, and the industrial schedule must be assessed together.

6. Queues measure pressure, not delivered capacity

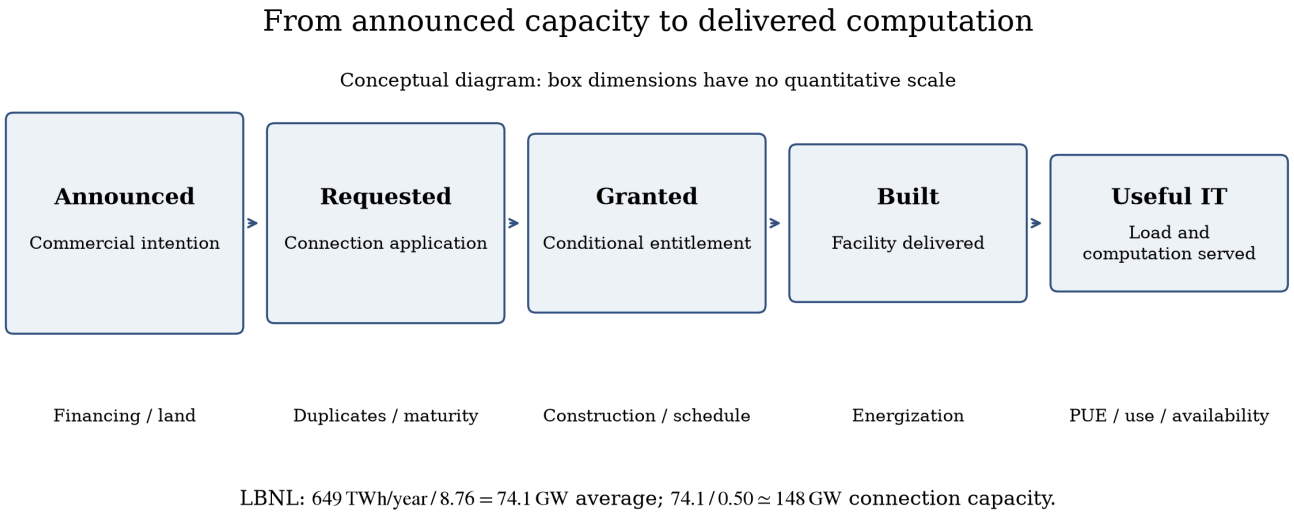
The IEA identifies more than 2,500 GW awaiting connection worldwide, combining generation, storage, and large loads. Indicative grid infrastructure lead times reach 5–15 years, compared with 1–3 years for data centers. Procurement creates another bottleneck: an IEA survey reports 2–3 years for cables and up to four years for large transformers. These indicative durations do not replace a supplier-specific schedule. IEA, Grids; IEA, equipment.

The US generation queues tracked by LBNL contained approximately 1,312 GW of generation and 749 GW of storage at the end of 2025. For projects completed in 2025, the median time between application and operation exceeded five years. These are transmission-connected generation applications, not the queue of computing loads. LBNL, Queued Up 2026.

An appropriate representation tracks each project i by state and date:

$$K_{\text{served}}(t) = \sum_i K_i \, z_i(t) \, r_i(t),$$

where z_i indicates actual commissioning and r_i its ramp-up. A weighted sum $\sum_i p_i K_i$ can support planning, but the probabilities are neither independent nor directly observed: several applications may involve the same customer, financing source, or scarce equipment. A funnel becomes quantitative only after deduplication and documentary verification.



F05. Author-generated schematic with no assumed volumes or conversion rates. The categories are not a strictly nested statistical cohort: multiple applications, expansions, cancellations, and partial commissioning must be reconciled project by project. IT MW do not directly measure useful FLOP. The LBNL 2025 Update example, published June 2026, converts 649 TWh/year into 74.1 GW of average power, then approximately 148 GW of connection capacity under an assumed electrical load factor of 50%. This factor is neither GPU utilization nor available firm capacity. Source: source

7. Size a firm power supply

An annual energy balance is a necessary but insufficient condition:

$$\sum_s \int_{\mathcal{T}} P_s(t) \, dt \geq E_{\text{load}} + E_{\text{losses}}.$$

The operating constraint must also be satisfied at every time step, accounting for outages and transmission limits:

$$P_{\text{grid}} + P_{\text{dispatchable}} + P_{\text{variable}} + P_{\text{discharge}} - P_{\text{storage charging}} \geq P_{\text{DC}} - P_{\text{curtailable}}.$$

Storage adds an intertemporal constraint:

$$S_{t+1} = S_t + \eta_c P_{c,t} \Delta t - \frac{P_{d,t} \Delta t}{\eta_d}, \qquad 0 \le S_t \le S_{\max}.$$

Its maximum power does not determine its endurance. For a complete supply interruption lasting τ at constant load P_L, the minimum useful energy is $P_L \tau$, before margins, aging, and conversion losses. The governing duration may be meteorological, logistical, or technical; choosing it to obtain an artificially lightweight battery would invalidate the comparison.

Reliability must be expressed at the service level, for example through expected energy not served:

$$\mathrm{EENS} = \mathbb{E}\left[\int_{\mathcal{T}} \max\left(0, P_{\mathrm{load}} - P_{\mathrm{available}}\right) dt\right].$$

The target depends on the contract. Interruptions to a curtailable training workload do not have the same consequences as interruptions to interactive inference, but restarts, checkpoints, and deadlines still carry costs. Annual average availability conceals long and correlated interruptions. Redundancy does not eliminate common-mode failures in fuel supply, substations, or connections.

8. Investment implications

A 2029–2030 decision must value computation that can actually be sold after energization. A physical schedule indicator is

$$C_{\mathrm{delivered}} = \int_{t_{\mathrm{service}}}^{t_{\mathrm{end}}} \eta_{\mathrm{eff}}(t) P_{\mathrm{fac}}(t)\, dt.$$

Two sites with the same final power rating can deliver very different volumes if one opens several years later. The best trajectory therefore selects sites whose connections, equipment, and commercial commitments are consistent, then installs accelerators in step with secured demand. Flexibility can accelerate certain connections; it must be sold and priced as a service property, without converting interruptible supply into a promise of continuous service.

Model validation must begin by reconciling system boundaries and revisiting available observations. For each new publication, retain the previous forecast, identify modified parameters, and separate historical revisions, methodological changes, and prospective revisions. Growth in consumption attributed to accelerators should not be reintroduced a second time through an independent campus extrapolation. Similarly, hardware replacement can improve efficiency while increasing installed power if the number of machines grows; both effects must appear in the cohort balance.

For a concrete project, the decisive checkpoints are the contractual connection date, capacity actually reserved, orders for long-lead equipment, availability of computing equipment, and commercial ramp-up. Quarterly monitoring contributes more to a financing decision than multiplying global scenarios without local observation. Projections identify pressure and test robustness; they do not replace evidence of implementation.

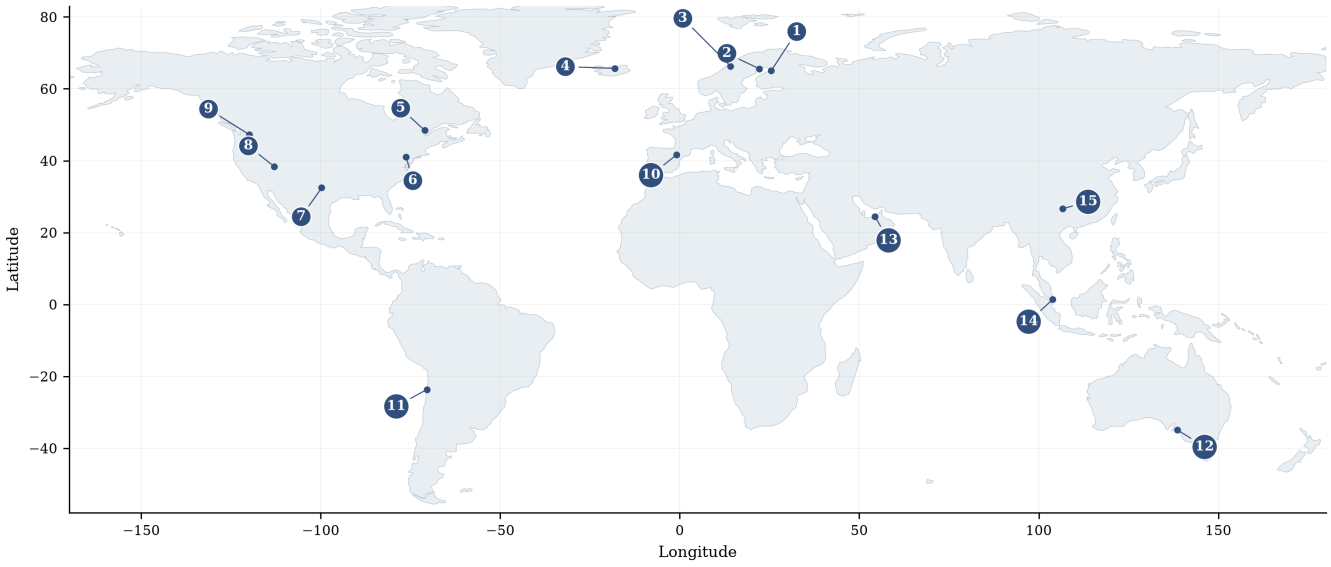
None of the forecasts examined justifies treating terrestrial deployment as energetically untenable. They justify more demanding selection of grid nodes, rigorous industrial planning, and multiple consumption scenarios. An airborne or orbital architecture remains a relevant prospective option if it demonstrates, for an equivalent service, firm supply, a credible schedule, and a lower complete computing cost than accessible terrestrial alternatives, including maintenance. The grid constraint then becomes a measurable comparison parameter, rather than a sufficient argument for imposing a conclusion.

The geography of computation in 2029–2030: from candidate regions to deliverable capacity

The optimal geography of computation emerges where identifiable demand meets deliverable infrastructure. It cannot be inferred from latitude, average spot prices, or a country's renewable energy volume. For commissioning in 2029–2030, the decisive variable is power accessible at a specific substation, on a date compatible with customer commitments, at an acceptable total cost and availability. Proximity to generation, climate, and connectivity then become optimization factors. This hierarchy is particularly important for computing equipment whose economic value can decline before a delayed connection is even completed.

The atlas selects fifteen representative regions, each associated with a reference city. Coordinates are used only for mapping and distance calculations. They indicate neither reserved land, available substations, nor projects that have secured permits. Northern Finland, for example, is a credible prospecting priority because its grid operator describes connection opportunities; this does not demonstrate that a specific parcel can accommodate several hundred megawatts. The same distinction applies to US nuclear power plants, Canadian hydropower, and new geothermal resources.

F06 | Fifteen regions to investigate for 2029–2030



01 Northern Finland — Oulu
02 Northern Sweden — Luleå
03 Helgeland, Norway — Mo i Rana
04 Northern Iceland — Akureyri
05 Quebec — Saguenay
06 Pennsylvania/PJM — Berwick
07 Texas/ERCOT — Abilene
08 Utah geothermal — Milford
09 Washington/Columbia — Quincy
10 Aragon, Spain — Zaragoza
11 Atacama, Chile — Antofagasta
12 South Australia — Adelaide
13 Abu Dhabi, UAE — Abu Dhabi
14 Singapore — Singapore
15 Guizhou, China — Guiyang

Approximate city markers. No reserved land or MW. Numbers are identifiers, not scores.

Geographical markers, without quality scores or demonstrated land or power availability. Natural Earth 1:110m base map, public-domain data; see data/land.geojson. The Utah region also represents the Nevada geothermal option discussed in the text.

Method: distinguish resource, contract, and delivery

Three quantities must remain separate: installed generation capacity, capacity already under contract, and additional power deliverable to the project. A contract signed by another buyer demonstrates that a transaction is possible; it does not constitute available inventory. Similarly, a grid access permit does not certify that all necessary infrastructure has been completed. The physical requirement is written at the delivery point and for the specified critical states:

$$P_{\text{deliverable}}(s,t,\omega) \geq P_{\text{critical}}(t), \qquad t_{\text{completion}}(s) \leq t_{\text{customer commitment}},$$

where ω represents, among other conditions, an infrastructure outage, a weather event, or a generation constraint. An annual energy assessment cannot replace this power verification. Guaranteed capacity in a dry year, during a winter peak, or following a transmission line outage can differ substantially from rated capacity.

The documented facts in this atlas come from grid operators, public authorities, operators, or buyers. Their status is specified: completed development, agreement, program, first power delivery, or forecast. By contrast, the prices and lead times used in the figures are exploratory parameters. No price is attributed to a supplier, and no lead time is presented as an offer. Monetary ranges are expressed in constant 2026 USD; applying them to 2030 is an explicit assumption. Foreign exchange risk, local taxes, and contractual provisions require a separate assessment.

Northern Europe: a real advantage, contingent on grid nodes

1. Oulu and northern Finland. In August 2026, Fingrid reported that connection agreements covered nearly 5 GW of final project capacity for data centers, while identifying opportunities for rapid industrial development in the north. These two statements must be read together: the region is attractive, but demand is already substantial. Due diligence must address the substation, initial capacity, expansions, and ramp conditions. Cold weather favors dry cooling; it eliminates neither summer peaks nor freezing constraints. Telecommunications redundancy must be physical, rather than merely commercial: two operators may use the same corridor. Fingrid, August 2026.

2. Luleå–Boden and northern Sweden. Svenska kraftnät describes reinforcements that could enable approximately 2,000 MW of additional withdrawal capacity in the north between 2028 and 2030. This is a grid opportunity intended for several industrial uses, not an allocation reserved for computing. A credible proposal must identify the infrastructure on which it depends and verify capacity already allocated. Heat recovery markets can improve local economics, provided the temperature, seasonality, and price actually paid are documented. For a 2030 campus, an already secured connection may be more valuable than a better energy price that remains theoretical. Svenska kraftnät, northern development.

3. Mo i Rana and Helgeland, Norway. The refurbishment and expansion of the Rana substation targets completion in 2029, conditional on a spring 2027 start. Statnett also identifies an industrial flicker constraint that must be resolved before the additional capacity can be used. This case illustrates the difference between abundant hydropower and deliverable power quality. A former industrial site may be attractive, but its previous connection rights must not be assumed to be transferable or retained. Design must incorporate fiber corridors, potential coastal corrosion, and hydrological conditions affecting firm supply. Statnett, Rana substation.

4. Akureyri and northern Iceland. The PPA announced by Landsvirkjun with atNorth covers up to 12 MW of firm power for five years, effective in the second half of 2026. The operator attributes this opportunity to more favorable transmission conditions in the north than in the south. This is concrete evidence at the scale of tens of megawatts, not proof of immediate gigawatt scalability. FARICE-1, DANICE, and IRIS are existing cable infrastructure; the new AUÐUR cable, announced as being in preparation in April 2026, must remain a future option. Iceland is therefore a candidate for contractually secured increments, with a specific assessment of the cost and resilience of international connections. Landsvirkjun, Akureyri PPA, Farice, AUÐUR project.

North America: firm power does not eliminate competition for access

5. Saguenay and Quebec. Hydro-Québec's February 2026 tariff proposal envisages an average price of 13 Canadian cents per kWh for new large data centers. This source presents a proposal subject to regulatory proceedings; it is insufficient to confirm the final tariff applicable to a project. The assumption that Quebec invariably supplies power at historical industrial tariffs must therefore be rejected. Project selection, allocated power, and winter peaks are as important as annual hydropower volume. The climate favors dry cooling, but does not justify monetary comparisons that mix CAD and USD. Hydro-Québec, tariff proposal.

6. Berwick, Pennsylvania, and the PJM system. The Talen–AWS agreement announced in 2025 provides for up to 1,920 MW from Susquehanna, with full ramp-up no later than 2032 and delivery through the grid. This precedent demonstrates a contracting route around existing nuclear power, but that capacity is already committed. Meanwhile, PJM reports a 6,831 MW shortfall relative to its reliability requirement in the 2028–2029 capacity auction. A prospective site must therefore assess capacity charges, co-location rules, redundancy, and reinforcements. A 2030 plan based on an existing asset must be distinguished from a plan contingent on completing a new reactor. Talen, SEC-filed press release, PJM, adequacy and large loads.

7. Abilene and Texas. Large-load applications tracked by ERCOT reached approximately 410 GW in its March 2026 update, with a substantial share still lacking submitted studies. This figure is neither certain future demand nor deliverable power. An August 2026 notice subsequently postponed classification deadlines in the Batch Zero process. Assuming a rapid Texas connection by default is therefore indefensible. Configurations combining renewables, storage, the grid, or local gas remain candidates, subject to permits, equipment availability, and reserves. Extreme heat must be correlated with cooling consumption and electrical constraints. ERCOT, application status, ERCOT, Batch Zero notice.

8. Milford, Utah, and the Nevada geothermal option. The first power delivery announced at Cape Station on 24 September 2026, together with the 396 MW PPA announced with Google for a 2028 target, makes advanced geothermal an industrial option worth examining for 2030. These developments do not constitute unallocated power available to a new entrant. In Nevada, the 115 MW procurement arrangement announced by Google as approved confirms the value of coordination among generator, grid, and buyer. The benchmark must use net power, seasonal performance, pumping requirements, and reservoir evolution. The resource, fiber, and delivery point may occupy different locations. Fervo, Cape Station, Fervo, 396 MW PPA, Google, Nevada.

9. Quincy and the Columbia Basin. Grant PUD plans reinforcements to Quincy's grid for 2029, but its August 2026 resource plan anticipates a qualified capacity shortfall under the WRAP program as early as summer 2030. The contrast is decisive: strengthening transmission and guaranteeing generation adequacy are separate problems. An offer must specify supply beyond 2030, rather than only access to existing infrastructure. Because Quincy is inland, it would also be incorrect to assign it the mild Pacific coastal climate without calculation. Water, summer heat, and hydropower generation in dry years must be assessed locally. Grant PUD, Quincy grid, Grant PUD, 2026 IRP.

Solar regions and Asian markets: distinct configurations

10. Zaragoza and Aragon. Red Eléctrica's publication of demand access capacity by node enables a more precise assessment than national investment announcements. The 11.8 GW granted nationally since 2022, but not yet operational at the time of the February 2026 publication, is not unallocated capacity in Aragon. Solar and wind must be assessed against deficit hours, grid charges, and storage. The actual substation, summer temperature, and basin-specific water rights are the relevant selection parameters. Red Eléctrica, demand access capacity.

11. Antofagasta and coastal Atacama. Chile's PDATA 2024–2030 plan seeks, among other objectives, to decentralize the industry and reduce its energy and water impacts. The relevant comparison is

between a campus close to solar generation and a coastal campus supplied through transmission. Neither obtains nighttime power at no cost. The creation of Humboldt Connect in 2025 improves the prospects for trans-Pacific connectivity, but establishes neither that the cable has entered service nor that backhaul from Antofagasta to its landing points exists. Freshwater, site hazards, and any seawater intakes must be included before drawing cost conclusions. MinCiencia, PDATA, Subtel, Humboldt Connect.

12. Adelaide and South Australia. AEMO's 2026 Integrated System Plan selects a combination of renewables, networks, storage, and gas backup for the NEM system. This is a system trajectory, not an allocated connection in Adelaide. The campus must be assessed using local resources and its firming requirements. Distance from European and US markets may be acceptable for some deferrable processing, but detrimental to rapid interactions. NEM conclusions must not be transferred to Perth's separate WEM system. AEMO, 2026 ISP.

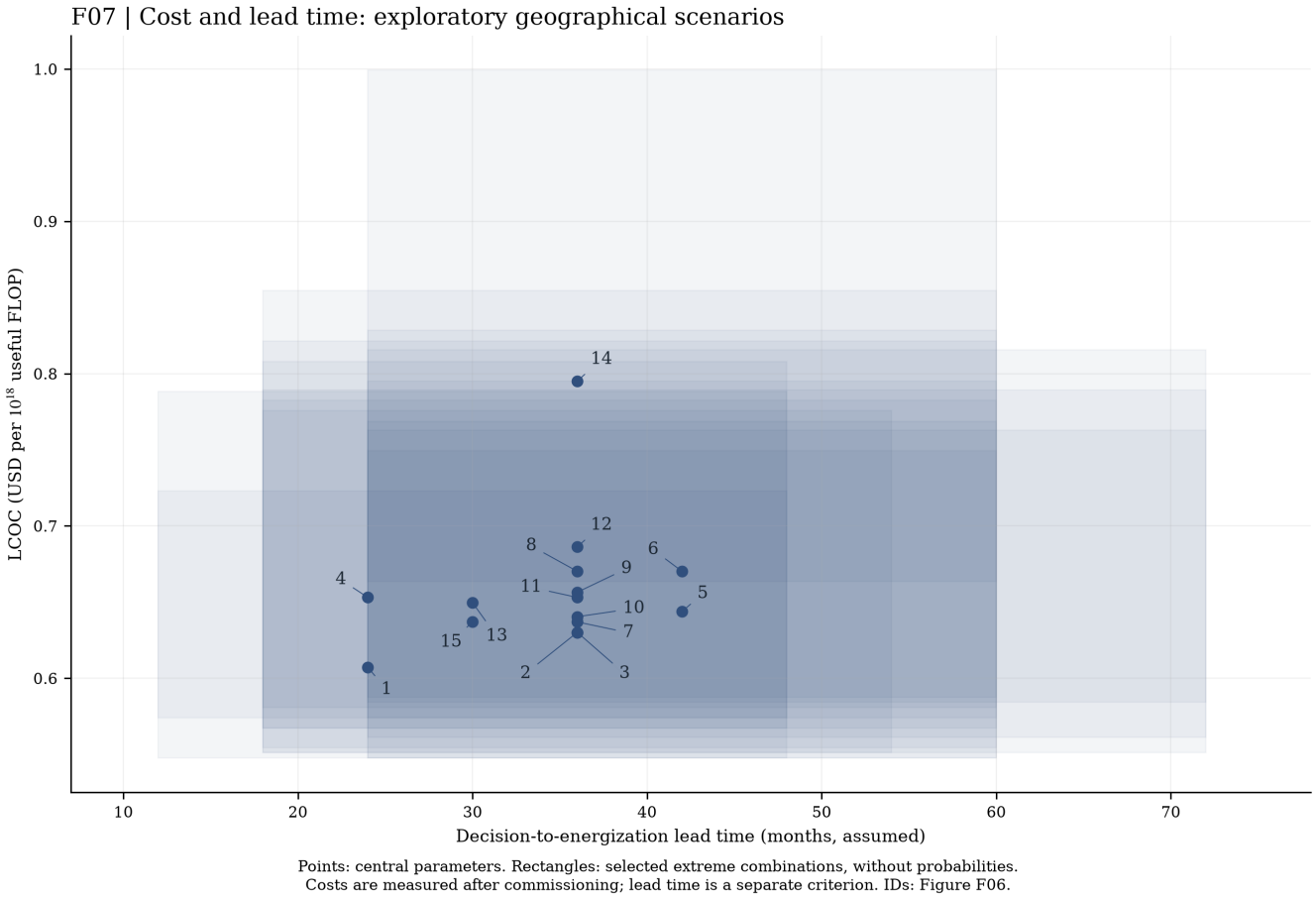
13. Abu Dhabi. The Masdar–EWEC project combines 5.2 GW of direct-current photovoltaics with 19 GWh of storage to target 1 GW of continuous output. It is an essential counterfactual to airborne solar: panels and batteries can be oversized on the ground under radically different mass constraints. The announcement and commencement of construction do not, however, replace guaranteed hourly performance. The balance must include dust, hot-cell efficiency, battery aging, and cooling consumption. Barakah nuclear generation is another possible system component, without being automatically available to the project. Masdar, solar and storage project.

14. Singapore. The IMDA roadmap targets at least 300 MW of additional near-term capacity, accompanied by efficiency improvements and green energy development. It confirms a policy of managed growth, not automatic access to capacity. For Asian inference and local data, proximity to demand can justify energy costs above those of a remote site. The hot and humid climate makes loop temperatures, liquid cooling, and water–electricity tradeoffs particularly important. Energy imports must be assessed through their own contracts and infrastructure. IMDA, Green Data Centre Roadmap.

15. Guiyang–Gui'an and Guizhou. In September 2026, the province joined a national computing interconnection mechanism coordinated with telecommunications operators and the power grid. This case merits comparison for Chinese demand, without assuming immediate interchangeability with international markets. Accelerator access, contracts, data, and interconnections must be verified for each workload. Plateau temperature is a design variable, not proof of decarbonized electricity. An announced capacity in FLOP cannot establish the additional firm power available. Guizhou government, interconnection node.

Price, schedule, and conditional frontiers

The figures use selected price, infrastructure, and lead-time ranges to explore sensitivity. Availability of 0.995 and PUE of 1.15 are held constant in this submodel to isolate these parameters. This convention does not claim that fifteen real sites would have identical thermal performance. Climate differences require an additional hourly simulation. The rectangles in Figure F07 represent extreme combinations without probabilities; their overlap does not define a statistical distribution.



Assumptions only: delivered tariffs, infrastructure, and lead times are neither local offers nor forecasts. Rectangles combine low/high prices and infrastructure costs; availability is fixed at 0.995 and PUE at 1.15. Costs and output cover six years from commissioning; no implicit economic delay penalty is applied. Overlaps are not statistical distributions. Rankings depend on the selected parameters.

Robust cost dominance would require, within the specified uncertainty set:

$$\max_{\theta \in \Theta_A} LCOC_A(\theta) < \min_{\phi \in \Theta_B} LCOC_B(\phi).$$

The present ranges overlap substantially; they therefore do not demonstrate a robust regional economic winner. Costs are calculated after commissioning. Lead time appears as a separate criterion: no fictitious revenue loss or automatic immobilization of GPUs is added. A complete commercial assessment would use the actual purchasing schedule, order book, and cost of delay.

For a campus actually consuming 100 MW of IT power continuously, a PUE of 1.15 yields 1.0074 TWh annually. A difference of 10 USD/MWh therefore changes the bill by 10.074 million USD per year. This sensitivity may exceed the savings obtained from a better climate. Conversely, electricity prices must not obscure the cost of losing a year under a computing contract.

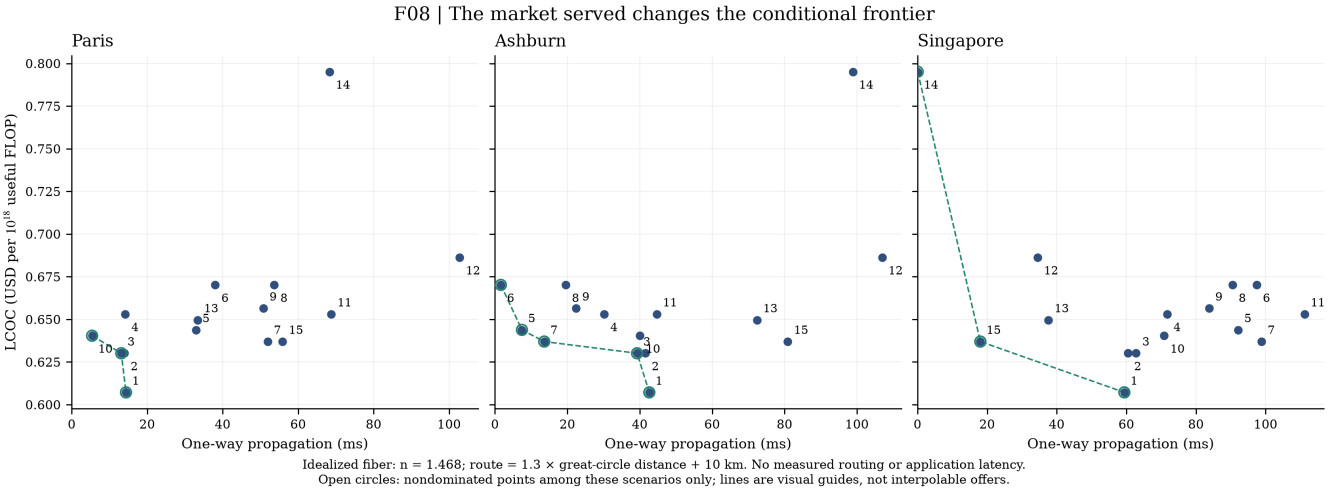
Paris, Ashburn, and Singapore: the customer shifts the optimum

Propagation is calculated here with a fiber group index of 1.468, a hypothetical 30% detour relative to the great-circle distance, and 10 km of local access:

$$\tau_{\text{one way}} = \frac{1.468\,[1.3 D_{\text{great circle}} + 10\,000]}{c}.$$

All lengths are in meters. This quantity is one-way optical propagation, not measured RTT or application latency. Existing cables, borders, landing points, and incidents may impose additional detours. Figure F08 nevertheless shows why an economic preference cannot be independent of the market served:

Singapore can remain relevant for local customers despite a high hypothetical central cost, whereas a Nordic region may suit European batch processing.



Calculated one-way optical propagation, not RTT or service latency. Hypothetical routes: 1.3 times the great-circle distance plus 10 km, with a group index of 1.468. Actual detours, cable landings, queuing, switching, and computation are excluded. Exact Pareto frontier for the fifteen hypothetical central costs, not a universal geographical optimum.

Latency between a user and a campus must be distinguished from internal training latency. Moving an entire cluster to an energy-rich region preserves short local interconnections; arbitrarily distributing synchronous training across several countries can instead substantially impair efficiency. Checkpoint migrations, data preparation, and independent jobs face different constraints. A scheduling policy must therefore choose the unit of relocation before comparing sites.

Cooling, water, and parcel-level due diligence

The qualitative F09 table accompanies the data without inventing a numerical water-stress index. It distinguishes cold conditions, dry heat, humid heat, and marine environments. A land-based coastal cooling system should be the first counterfactual for a submerged installation: it can benefit from seawater while preserving equipment accessibility, as Hamina illustrates. Offshore hulls, mooring, cables, and maintenance must justify their costs through demonstrated savings; cold water supplies no additional electrical energy. Google, Hamina.

The next phase is to request offers for comparable parcels: initial power and expansion, N-1 conditions, a binding date, complete price, hourly profiles, permits, dry- and wet-bulb temperatures, water rights, two fiber routes, and measured availability. Contractual electricity must be reconciled with physically delivered energy and its hourly carbon intensity. An increment sold to a creditworthy customer, energized on a firm date, and capable of expansion may be more valuable than a theoretically superior project dependent on several unfinished infrastructure assets. Converting a regional advantage into a deliverable service is the actual siting problem for 2030.

Thermodynamics: ambient temperature is not sufficient

From computing power to heat-transfer area

Almost all electrical energy consumed within the IT boundary ultimately becomes heat. The fraction exported through the useful signal is negligible relative to a MW of IT power. A platform may reject some photovoltaic and propulsion losses locally, but this separation must be physical and documented. It does not justify arbitrarily removing pumps, converters or storage losses from the energy balance.

For a surface at temperature T_s with equivalent thermal capacitance C_{th}:

$$C_{th}\frac{dT_s}{dt} = \dot{Q}_{IT} + \dot{Q}_{aux} + \dot{Q}_{abs} - hA(T_s - T_\infty) - \varepsilon\sigma A(T_s^4 - T_{rad}^4).$$

The coefficient h is not a constant determined by altitude. It depends on the fluid, flow regime, geometry and relative velocity. Cold, low-density stratospheric air increases the temperature difference but reduces convective transfer for comparable geometry and velocity. For a smooth, isothermal flat plate in laminar flow:

$$Re_L = \frac{\rho U L}{\mu}, \qquad Pr = \frac{\mu c_p}{k}, \qquad \overline{Nu}_L = 0.664 Re_L^{1/2} Pr^{1/3}, \qquad \overline{h} = \frac{k\overline{Nu}_L}{L}.$$

This correlation must be used within its domain of validity; a large envelope, a fin array or separated flow requires a different model. The SACP calculation uses a plate characteristic length of 1 m, distinct from the vehicle length. At 22 km, $T_s = 350$ K and $U = 10$ m/s, it yields approximately $h = 2.48$ W/m²/K and 326 W/m² of convection. An effective radiative temperature of 250 K adds approximately 566 W/m², giving 893 W/m² in total. This radiative background is an assumption; ambient air at 218 K is not a black enclosure surrounding the machine. Lienhard, flat-plate heat-transfer correlations.

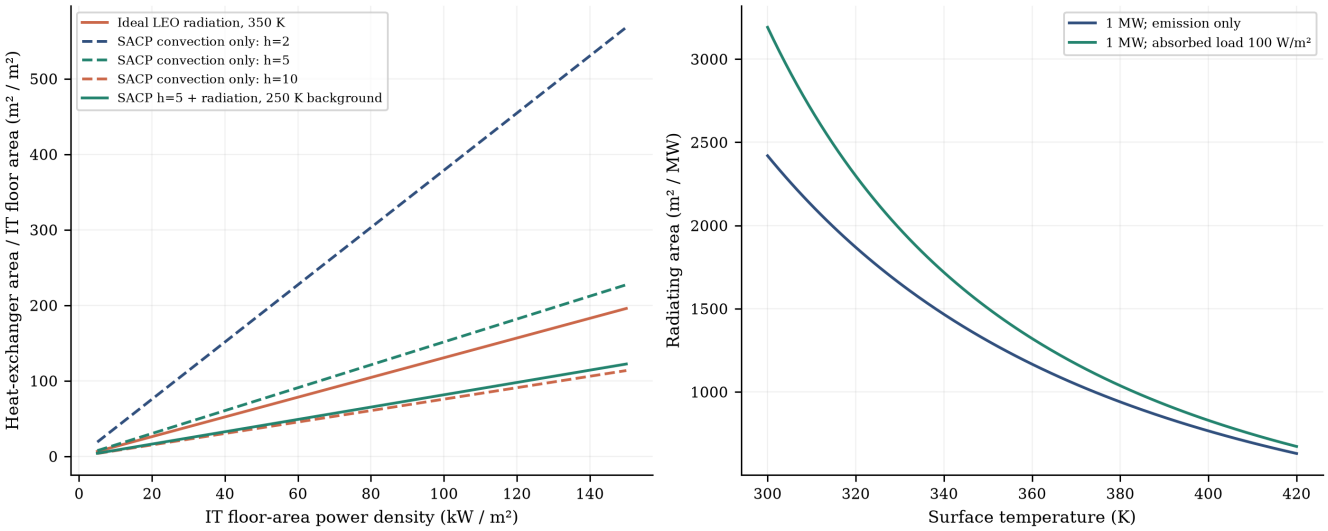
In vacuum, the ideal limit becomes

$$q''_{rad} = \varepsilon\sigma(T_s^4 - T_{space}^4), \qquad \sigma = 5.670374419 \times 10^{-8}\ \mathrm{W\,m^{-2}K^{-4}}.$$

At 350 K, an ideal blackbody emits approximately 851 W/m²; a surface with emissivity 0.9 emits 766 W/m². Rejecting 1 MW therefore requires approximately 1,306 m² of emitting area in this ideal case. Absorbed sunlight, albedo, terrestrial infrared radiation and view factors reduce the net capacity. The central orbital model assumes 100 W/m² of absorbed radiation, giving a net 666 W/m². This parameter does not come from a detailed orbital simulation. NASA, small-satellite thermal control.

Under the SACP convention, the radiative background at 250 K already includes environmental infrared radiation; no residual solar or albedo input is added, implying effective shading. Under the LEO convention, the 100 W/m² instead aggregates external inputs. The same infrared radiation must not be counted twice. At the central parameter values, an additional 126.7 W/m² of absorbed sunlight would eliminate the SACP heat-rejection advantage over ideal LEO, corresponding to an absorptivity of 0.0975 at normal incidence under 1,300 W/m². This conditional threshold illustrates sensitivity to integration and coatings.

SACP convection alone must satisfy $h > 766/(350 - 218.6) \approx 5.83$ W/m²/K to exceed the ideal orbital radiator at the same temperature. The 10 m/s flat-plate case remains below this threshold; combined convection and radiation can exceed it. An intrinsic thermodynamic superiority of the stratosphere therefore cannot be inferred from “−55 °C” alone. Optimization must address the complete system, including mass and auxiliaries.



F16. Consistent normalization of heat-transfer area to rack power density. Left: h is an imposed sensitivity parameter, not a property determined by air at −55 °C alone. Right: temperature, emissivity and absorbed load determine the required area in vacuum; each emitting surface is counted once.

Figure F16 resolves a common ambiguity: at fixed total power, increasing rack density does not automatically change the required area. If $d_{IT} = P_{IT}/A_{IT}$ and q''_{net} is the allowable heat flux:

$$\frac{A_{rejection}}{A_{IT}} = \frac{d_{IT}}{q''_{net}}.$$

The vertical axis therefore represents an area ratio. Comparing densities reveals the necessary separation between IT compactness and heat-rejection area. The right-hand panel shows another design variable: increasing temperature strongly improves emission, but only if the machine tolerates the corresponding internal temperatures.

Junction temperature, pumping and heat recovery

For an aggregate thermal path with resistance R_{th},

$$T_j = T_r + R_{th}\dot{Q}, \qquad T_j \le T_{j,max}.$$

This notation represents a resistance appropriately referred to the system boundary under study; the entire MW does not pass through a single chip interface. The calculation must be distributed across modules, hot spots verified and cooling flow accounted for. A radiator at 350 K requires a hotter loop and hotter junctions. Figures sweeping 320–400 K hold accelerator efficiency constant to isolate the area effect; they do not demonstrate that this entire range is usable with a commercial accelerator.

If a heat pump raises the heat-rejection temperature, it consumes work. Its Carnot limit in refrigeration mode is

$$COP_{max} = \frac{T_c}{T_h - T_c}, \qquad W \ge \frac{\dot{Q}_c}{COP_{max}}, \qquad \dot{Q}_h = \dot{Q}_c + W.$$

The radiator may shrink, while the generator, compressor and heat to be rejected increase. An optimum exists only after these loops have been closed. Similarly, fans and pumps require power $P = \Delta p\dot{V}/\eta$, which must be included among auxiliaries. Maximizing h without accounting for this energy yields a biased comparison.

Heat exported on the ground may have commercial value, but it should be deducted from costs only when there is a buyer, a useful temperature and seasonal alignment. A cold city may lack a district-heating network accessible from the site; a uniform annual revenue would then overestimate the benefit. The heat contract and connection expenditure must be included in the same model as the compute contract.

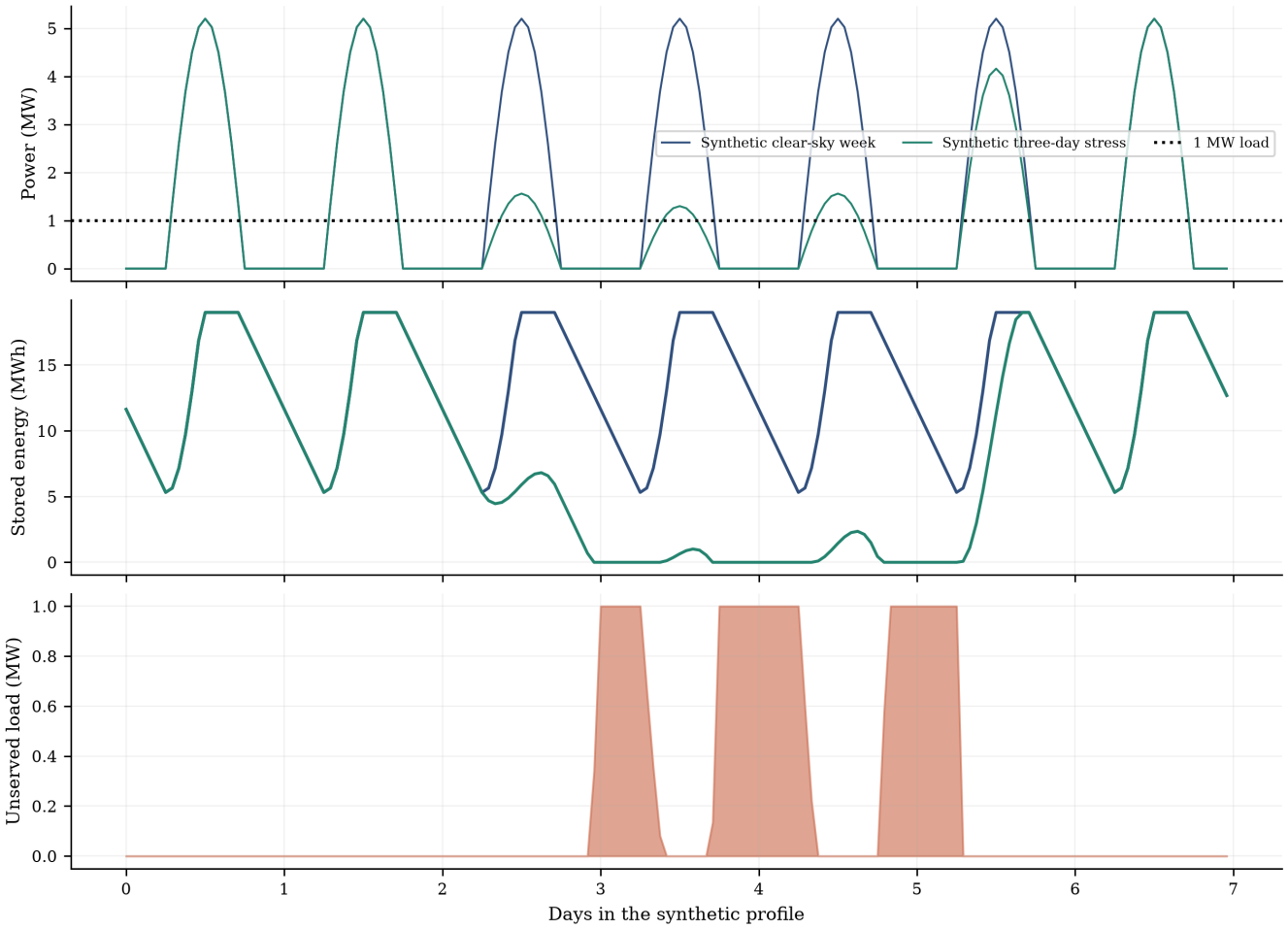
Irradiance and energy actually available

The reference value near 1,361 W/m² describes solar flux at Earth's distance, outside the atmosphere, on a surface normal to the radiation. It does not increase by 35–40% when ascending into the stratosphere. Such a percentage approximately compares this reference with 1,000 W/m² under certain terrestrial conditions; it compares neither meteorological years nor identical orientation profiles. The stratosphere still has atmospheric absorption and night. NASA, solar irradiance.

The appropriate energy integral is

$$E_{PV} = \int A_{PV}\eta_{PV}(T, age)G_{plane}(t)\eta_{elec}(t)\,dt.$$

The profile G_{plane} incorporates incidence, orientation, season and shading. Near-continuous orbital solar exposure can be a substantial advantage; it must be established along the actual trajectory. On the ground and in the stratosphere, large annual production does not eliminate hourly deficits.



F23. Deliberately synthetic hourly example: 5.2 MWp solar, 19 MWh battery, 1 MW load, charge/discharge efficiency 0.95, power limits 5/2 MW, and a 1 h time step. The state converges after repeated weeks. This stress test represents neither Abu Dhabi weather nor guaranteed performance of an actual project.

Stress test F23 uses a 1 MW load, 5.2 MW of peak photovoltaic capacity and 19 MWh of storage. Its profiles are entirely synthetic. During the adverse week, solar generation produces 183.7 MWh against 168 MWh of demand, yet approximately 34.0 MWh remains unserved under the selected battery model. Chronology and storage limitations explain this gap. This is not a reliability estimate for a real project: no Abu Dhabi meteorological year was simulated. It is a calculated counterexample to the inference that “sufficient total energy implies firm power.”

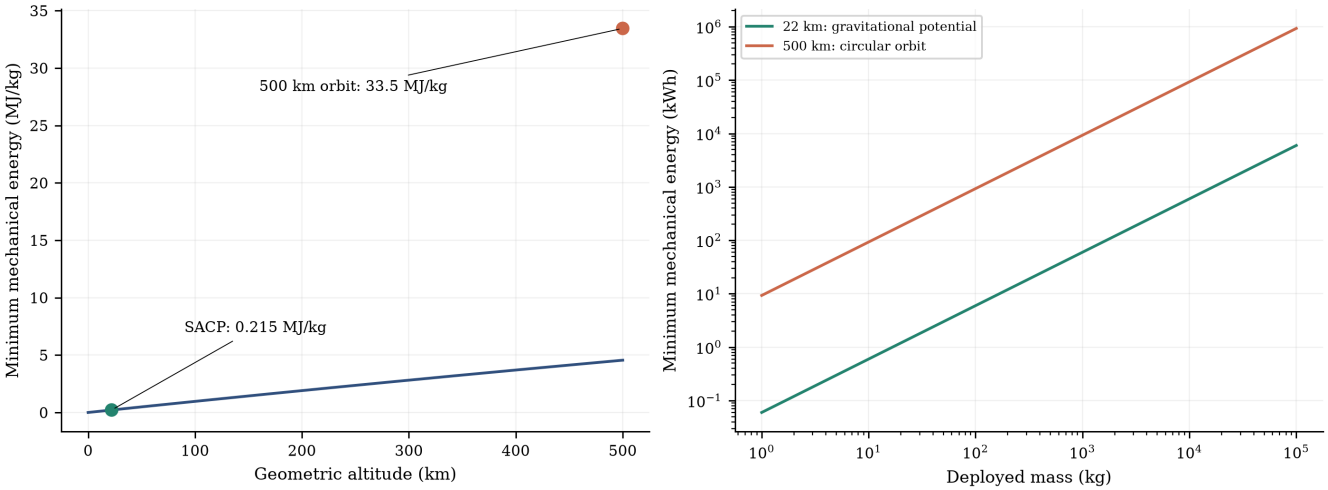
Deployment and communications: distance has several meanings

Raising a mass and placing it in orbit

The difference in potential energy between the surface and altitude h is

$$\Delta e_{pot} = \mu\left(\frac{1}{R_E} - \frac{1}{R_E + h}\right).$$

A circular orbit also requires a tangential velocity $v_c = \sqrt{\mu/(R_E + h)}$. At 500 km, the specific energy difference from rest at the surface is approximately 33.56 MJ/kg. Crediting a favorable terrestrial rotation of 465 m/s reduces it to approximately 33.45 MJ/kg. Figure F18 uses this second convention, whereas the rounded comparisons in the orbital chapter explicitly use the first. At 22 km, elevation alone represents approximately 0.215 MJ/kg.



F18. Mechanical lower bounds, not fuel consumption or cost. Aerostat ascent draws on buoyancy and atmospheric work; an equivalent speed of √(2gz) is not a propulsive delta-v. A favorable Earth-rotation contribution of 465 m/s is credited in the orbital calculation.

For two circular orbits with radii r_1 and r_2, the ideal Hohmann transfer is

$$\Delta v_1 = \sqrt{\frac{\mu}{r_1}}\left(\sqrt{\frac{2r_2}{r_1 + r_2}} - 1\right),$$

$$\Delta v_2 = \sqrt{\frac{\mu}{r_2}}\left(1 - \sqrt{\frac{2r_1}{r_1 + r_2}}\right), \qquad \Delta v_H = \Delta v_1 + \Delta v_2.$$

Setting $r_1 = R_E$ does not model an object initially at rest on the ground: this formula already assumes an initial circular orbit. Actual ascent requires a velocity budget including gravitational and aerodynamic losses, rotation, guidance and recovery. The Tsiolkovsky relation, $\Delta v = I_{sp} g_0 \ln(m_0/m_f)$, converts a propulsion budget into a mass ratio under its own assumptions; it does not by itself yield a delivered price.

An aerostat does not need to acquire orbital velocity. Its ascent involves buoyancy work exchanged with the atmosphere; $\sqrt{2gh}$ would merely be an energy-equivalent velocity. The electrical requirements of station keeping and operational life may dominate the initial lifting energy. Figure F18 must therefore not be interpreted as an LCOC reduction factor.

Optical link budget at 1,550 nm

For circular transmitting and receiving apertures in the far field, with aperture efficiencies η_t and η_r, the idealized gains are

$$G_t = \eta_t\left(\frac{\pi D_t}{\lambda}\right)^2, \qquad G_r = \eta_r\left(\frac{\pi D_r}{\lambda}\right)^2.$$

The link equation is

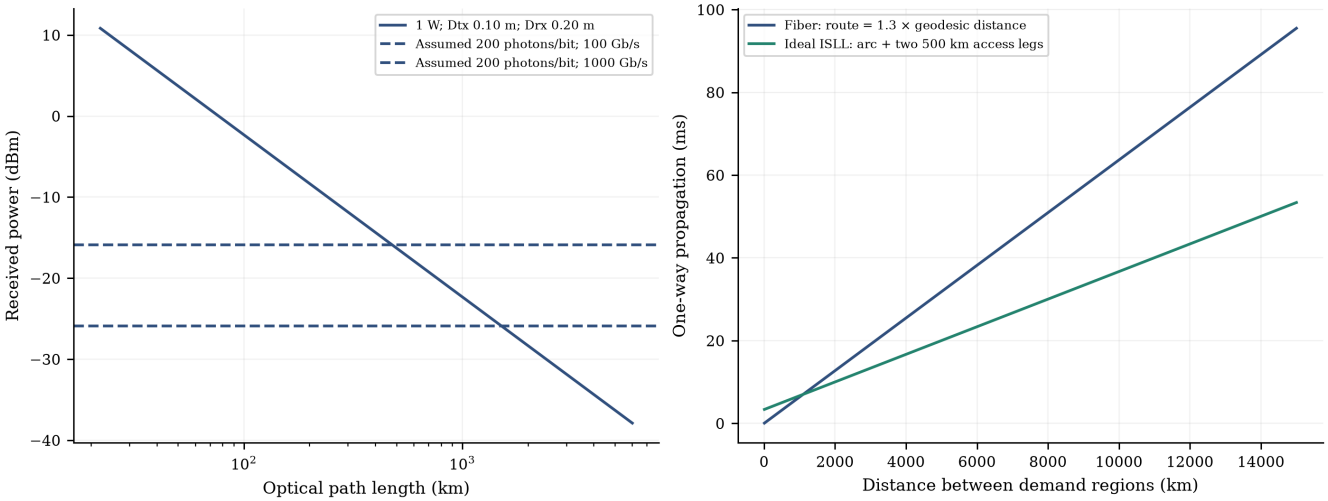
$$P_r = P_t G_t G_r \left(\frac{\lambda}{4\pi R}\right)^2 L = P_t \eta_t \eta_r \left(\frac{\pi D_t D_r}{4\lambda R}\right)^2 L.$$

Loss factors satisfy $0 < L \leq 1$. In decibels, gains are added, while free-space loss $20\log_{10}(4\pi R/\lambda)$ and additional losses are subtracted. Real optics require a beam profile, pointing, aberration control, a detector and a modulation chain. The preceding expression remains a geometric coupling approximation, not a certification of data rate.

Figure F17 uses 1 W of optical power, diameters of 0.10 and 0.20 m, aperture efficiencies of 0.60 and eight dB of assumed additional losses. The mean number of photons received per bit is

$$n_\gamma = \frac{P_r \lambda}{h_P c R_b}.$$

The 200-photon-per-bit lines serve only to visualize the effect of data rate, from 100 Gb/s to 1 Tb/s. A real receiver may require another value depending on modulation, noise, quantum efficiency, error correction and the allowable bit error rate. No operational data rate is inferred merely from crossing this line.



F17. Geometric optical link budget at 1550 nm and path comparison. The 8 dB allowance combines assumed pointing, optical and clear-sky losses. The photons/bit threshold is a modeling assumption, not a qualified receiver sensitivity. The ISLL path is an illustrative bound excluding processing, handover and actual orbital routing detours.

For a Gaussian beam, an angular error can be represented approximately by $L_p = \exp[-2(\theta_e/\theta_b)^2]$. As divergence decreases, the margin gained from a large aperture comes with more demanding pointing requirements. Turbulence and clouds impair ground-to-air and ground-to-space links. Inter-satellite links avoid clouds, but a service intended for a terrestrial user must still manage its ground access. Geographically diverse gateways reduce interruptions if their weather and networks are not excessively correlated. Availability computed by multiplying independent probabilities would not be justified without these data.

Propagation, refraction and application latency

The geometric propagation time through a medium of group refractive index n_g is

$$\tau = \frac{1}{c}\int_\Gamma n_g(s)\, ds.$$

Vertically, 22 km corresponds to 0.0734 ms in vacuum and 500 km to 1.668 ms. These are one-way paths to onboard computation. A round-trip time adds the return path; a relay to another terrestrial point adds another downlink. These paths must not be conflated. Atmospheric refraction modifies both the integral

and the trajectory; its effect must be calculated using pressure, temperature, humidity, wavelength and elevation. Applying a sea-level refractive index uniformly across 22 km overestimates the additional delay. At the order of magnitude discussed here, slant range, terrestrial fiber and processing generally dominate the refractive correction.

In the maps, fiber uses an assumed refractive index of 1.468 and a route length of $1.3d + 10$ km, where d is the great-circle distance. This is a reproducible assumption, not an observed commercial route. Space can gain an advantage in long-distance propagation because light travels faster in vacuum than in fiber; the two orbital access legs and routing nevertheless add a fixed length. In the idealized comparison of F17:

$$1.468 \times 1.3\,d > d + 1000$$

implies a crossover at $d > 1\,100$ km. This threshold is specific to the hypothetical route consisting of "a geodesic arc plus two 500 km access legs"; it does not predict the ping of a real constellation. Orbital planes, handovers, available routes and routers can shift it substantially.

Complete service latency is

$$L_{service} = L_{access} + L_{prop} + L_{routing} + L_{queue} + L_{execution} + L_{return}.$$

A platform directly above the client may have low propagation time and substantial queueing. For interactive inference, the distribution of delays and peaks is more useful than an average. For batch processing, initial latency may be secondary to cost and overall completion time.

Distributed training: the client link is not the GPU fabric

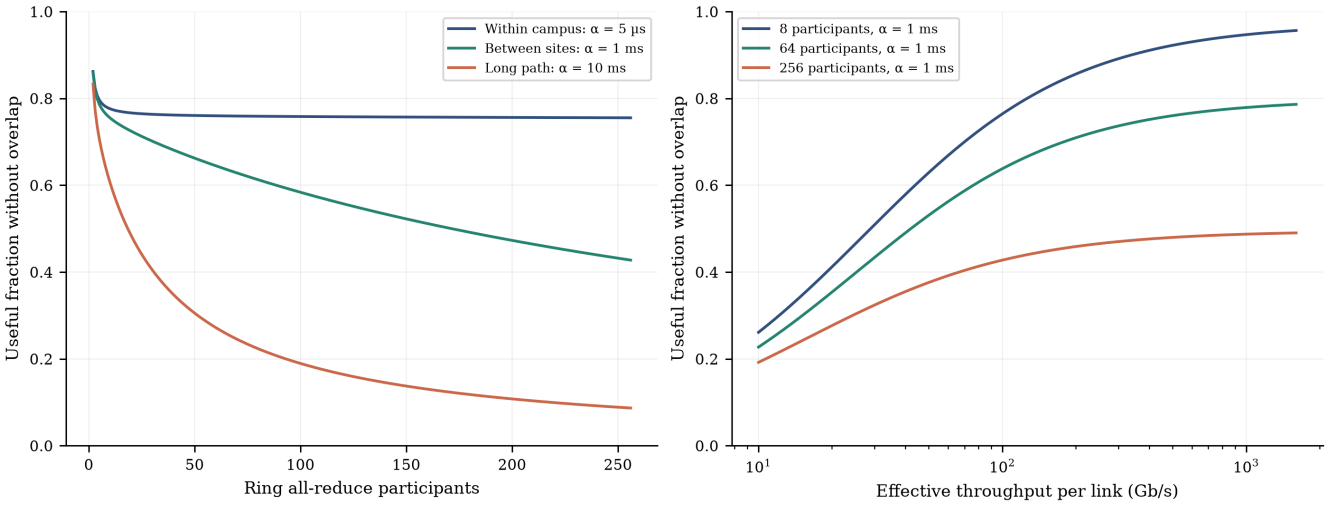
The efficiency of a set of nodes can be approximated by

$$\chi = \frac{T_{compute}}{T_{compute} + T_{communication} + T_{waiting} + T_{recovery}}.$$

For a simplified ring all-reduce with p participants, a message of size S, a per-step latency α and effective bandwidth B:

$$T_{AR} \simeq 2(p-1)\alpha + \frac{2(p-1)}{p}\frac{S}{B}.$$

This expression models two phases and assumes homogeneous links, without congestion or overlap. It isolates growth in the latency term; a real network may use trees, a hierarchy or other collectives. High bandwidth does not necessarily eliminate the cost of sequential steps.



F24. Alpha–beta model of ring all-reduce, without compute/communication overlap: 1 GB exchanged, 0.5 s of local computation, and 100 Gb/s on the left. Illustrative scenario, not a GPU benchmark or an AI model forecast. Topology, compression, pipelining, and hierarchical collectives may substantially change these curves.

Scenario F24 uses one GB, 0.5 seconds of local computation and 100 Gb/s links. It is not a benchmark of a commercial model. Its architectural implication is actionable: concentrate the most frequent exchanges within a campus or compact formation, and distribute tasks that are sufficiently independent. An entire constellation can sell substantial compute without constituting a single efficient cluster. The cost model uses $\chi = 1$ as its reference; any actual reduction increases LCOC approximately in proportion to $1/\chi$, at unchanged costs.

SACP in 2029–2030: close the balances before comparing costs

A stratospheric aerostatic compute platform can decouple its power supply from the connection of a terrestrial campus to the grid. This property warrants experimental research. It does not, however, demonstrate capital savings, industrial availability or gigawatt-scale deployment capability. The relevant comparison concerns a delivered computing service: useful computing capacity, availability, hardware replacement and discounted cost. The 2029–2030 horizon must therefore be studied as a window for demonstration and qualification, with explicit conditions for scaling. The model presented here simultaneously closes the lift, propulsion, solar generation, storage and heat-rejection balances. It is a reproducible exploratory calculation, not a certified preliminary design.

Reference atmosphere and geometry

The selected geometric altitude is $z = 22\,000$ m. The standard atmosphere provides a reference state, distinct from a local weather forecast. After conversion to geopotential altitude $H = R_g z/(R_g + z)$, with $R_g = 6\,356\,766$ m, the resulting values are $T_a = 218.574$ K, $p = 4.0475$ kPa and $\rho_a = 0.06451\ \mathrm{kg\,m^{-3}}$. These values follow the structure of the U.S. Standard Atmosphere, 1976. For an ideal gas in pressure and temperature equilibrium with the external air,

$$\rho_g = \frac{p}{R_g^{\mathrm{gas}} T_g}, \qquad m_{\text{excluding gas}} = (\rho_a - \rho_g)V.$$

For helium, $R_{\mathrm{He}} = 2077.264\ \mathrm{J\,kg^{-1}\,K^{-1}}$, giving $\rho_{\mathrm{He}} = 0.008914\ \mathrm{kg\,m^{-3}}$ and net mass lift of only $0.05560\ \mathrm{kg\,m^{-3}}$. Each additional kilogram therefore requires approximately $18\ \mathrm{m^3}$ of volume, before structural feedback. Low temperature does not compensate for the thinning atmosphere.

The force equilibrium explicitly includes buoyancy and weight, before cancellation of local gravity:

$$F_B = \rho_a V g, \qquad F_W = (m_g + m_{\text{excluding gas}})g, \qquad F_B = F_W.$$

The geometry is a prolate spheroid of length $L = 3D$, with semi-axes $a = 3D/2$ and $b = D/2$. With $e = \sqrt{1 - b^2/a^2}$,

$$V = \frac{4\pi a b^2}{3}, \qquad S = 2\pi b^2\left[1 + \frac{a}{be}\arcsin(e)\right], \qquad A_f = \pi b^2.$$

Lift equilibrium is insufficient: the geometry must also satisfy membrane, stability, maneuvering and launch constraints. The assumed envelope areal mass, $0.5\ \mathrm{kg\,m^{-2}}$, is an optimistic extrapolation. It is not a demonstrated performance for a hull of this size.

Coupled energy, propulsion and storage loop

The calculation considers $1\ \mathrm{MW}$ of IT at full load and $150\ \mathrm{kW}$ of auxiliaries. Electrical station-keeping power is represented by

$$P_{\mathrm{stat}} = \frac{\rho_a C_D A_f U^3}{2\eta_p}, \qquad P_L = 1.15 \times 10^6 + P_{\mathrm{stat}},$$

with $C_D = 0.05$, $\eta_p = 0.70$ and $U = 10\ \mathrm{m\,s^{-1}}$ in the central scenario. These coefficients do not come from tests of an equipped platform. U is the air-relative velocity required for the chosen trajectory; it does not

represent a guaranteed climatological mean. Panels, heat exchangers, suspension lines and control surfaces can simultaneously affect drag and stability.

For $H_d = H_n = 12$ h, equivalent irradiation $I_* H_* = 1300 \times 6$ Wh m^{-2}, and photovoltaic and electrical efficiencies of 0.30 and 0.90, the required area is

$$A_{\mathrm{PV}} = \frac{P_L (H_d + H_n/\eta_{\mathrm{rt}})}{I_* H_* \eta_{\mathrm{PV}} \eta_{\mathrm{elec}}}, \qquad \eta_{\mathrm{rt}} = 0.85.$$

Six equivalent hours of generation do not mean six hours of daylight. For each latitude and season, the integral must be replaced by a simulation of orientation, shading, cell temperature and irradiance. Ascent creates no increase in the “solar constant”: it reduces some atmospheric losses, while extraterrestrial irradiance remains close to 1361 W m^{-2} at one astronomical unit, as explained by NASA.

With an allowable depth of discharge $d = 0.80$, discharge efficiency $\eta_d = 0.95$ and system specific energy e_b,

$$E_{b,\,\mathrm{nom}} = \frac{P_L H_n}{\eta_d d}, \qquad m_b = \frac{E_{b,\,\mathrm{nom}}}{e_b}.$$

Hours and watts yield energy in Wh here. For $e_b = 250$ Wh kg^{-1}, storage reaches 18.60 MWh and 74.38 t. The battery is therefore a major design driver, rather than an accessory. Charge and discharge losses are added to the thermal balance; photovoltaic and propulsion losses are assumed to be rejected locally. This last assumption requires validation of the physical integration.

Cooling: cold ambient air does not determine the heat-transfer coefficient

The external temperature provides a large temperature difference, but the convective coefficient depends on density, velocity and characteristic length. For a laminar flat plate of length $\ell = 1$ m, with air properties evaluated at the film temperature $T_f = (T_s + T_a)/2$,

$$\mathrm{Re}_\ell = \frac{\rho_f U \ell}{\mu_f}, \qquad \overline{\mathrm{Nu}}_\ell = 0.664 \mathrm{Re}_\ell^{1/2} \mathrm{Pr}_f^{1/3}, \qquad h = \frac{k_f \overline{\mathrm{Nu}}_\ell}{\ell}.$$

At $T_s = 350$ K, the scenario yields $h \simeq 2.48$ W m^{-2} K^{-1} and $q''_{\mathrm{conv}} \simeq 326$ W m^{-2}. This plate provides a simplified heat-exchanger closure, not a correlation validated for an envelope hundreds of meters long. The length ℓ must remain explicitly distinct from L.

The general heat-flux balance is

$$q''_{\mathrm{net}} = h(T_s - T_a) + \varepsilon\sigma\left(T_s^4 - T_{\mathrm{rad,\,eff}}^4\right) - q''_{\mathrm{abs}}, \qquad A_{\mathrm{rad}} = \frac{\max(Q_d, Q_n)}{q''_{\mathrm{net}}}.$$

The central case uses $\varepsilon = 0.90$, $T_{\mathrm{rad,\,eff}} = 250$ K and $q''_{\mathrm{abs}} = 0$, implying effective solar protection. It gives $q''_{\mathrm{net}} = 892.5$ W m^{-2}, including 566.5 W m^{-2} of radiation, followed by $A_{\mathrm{rad}} = 1452$ m^2 after incorporating storage losses. Earth, the atmosphere, reflections and the Sun must enter an actual view-factor calculation. NASA's thermal review explains precisely why the radiative environment cannot be reduced to a vacuum temperature.

As a reference, an ideal, shaded orbital radiator viewing only a 3 K background emits 765.8 W m^{-2} at the same temperature and emissivity. The combined advantage of the SACP scenario is therefore conditional and moderate. It does not justify general thermodynamic superiority over terrestrial cooling, where dense forced convection can provide much higher coefficients.

Figure F16 in the thermodynamics chapter compares heat-rejection areas.

Central solution and decisive sensitivities

The numerical closure solves

$$(\rho_a - \rho_{\mathrm{He}}) V(D) = m_0 + \mu_E S(D) + \mu_{\mathrm{PV}} A_{\mathrm{PV}}(D) + m_b(D) + \mu_R A_{\mathrm{rad}}(D),$$

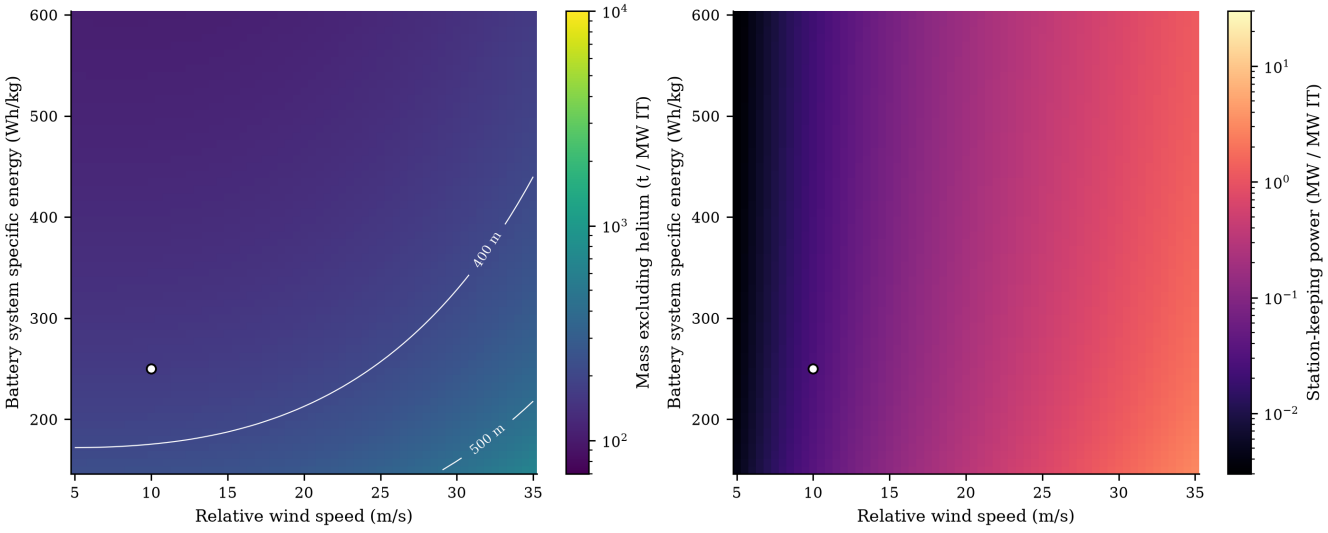
with $m_0 = 10$ t, $\mu_E = 0.5\ \mathrm{kg\,m^{-2}}$, $\mu_{\mathrm{PV}} = 1\ \mathrm{kg\,m^{-2}}$ and $\mu_R = 5\ \mathrm{kg\,m^{-2}}$. The solution is $L = 371.14$ m, $D = 123.71$ m and $V = 2.974 \times 10^6\ \mathrm{m^3}$.

Central quantity	**Calculated value**
Total mass excluding helium	165.35 t
Envelope	59.10 t
Battery	74.38 t
Photovoltaics	14.61 t; 14,605 m²
Heat exchangers	7.26 t; 1,452 m²
IT and fixed supports	10.00 t
Onboard helium	26.51 t
Station-keeping power	27.69 kW

Additional contingency, redundancy and qualification masses are not invented to close the model artificially: their absence must remain an explicit limitation. The 371 m result is an optimistic feasibility envelope, not a manufacturer's dimension.

Wind sensitivity is strongly nonlinear. Reclosing the entire vehicle while increasing U from 10 to 20, then $30\ \mathrm{m\,s^{-1}}$, raises the mass excluding helium to 185.1, then 259.1 t; propulsion reaches 0.239 and 1.009 MW, respectively. The additional power requires more solar generation and storage, enlarging the hull and further increasing its drag. Selecting a latitude therefore requires a wind time series and a movement policy, with service and airspace constraints.

The range 150–600 $\mathrm{Wh\,kg^{-1}}$ explores complete-system specific energy; it is not a technological forecast for 2030. At the two extremes, calculated total mass varies from 230.4 to 106.7 t, and length from 414.5 to 320.7 m. Cycling, containment, heating, end-of-life reserve and fault tolerance must be included in this system performance.



F14. Coupled closure at 22 km, with a continuous 1 MW IT load, a 12 h night and a 0.5 kg/m² envelope. White contours: vehicle length. First-order extrapolation; large vehicles are not structurally validated. Wind and storage are independent exploratory assumptions, not a climatology.

Structure and helium availability

Extrapolation at constant areal mass may underestimate large-scale structural constraints. For an illustrative spherical membrane,

$$\sigma_m = \frac{\Delta p\, r}{2t} \le \sigma_{\text{allow}} \quad \Rightarrow \quad t \ge \frac{\Delta p\, r}{2\sigma_{\text{allow}}}, \qquad m_{\text{membrane}} \propto r^3.$$

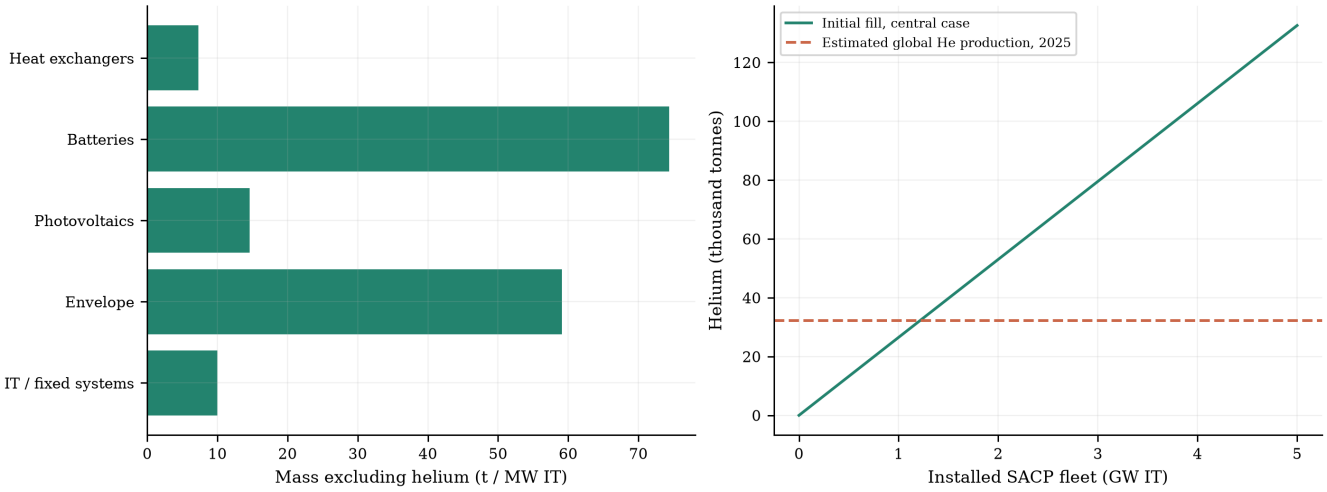
At constant overpressure and allowable stress, structural mass can scale like lift. The geometric economy of scale then partly disappears. An elongated hull requires its own analysis of stresses, seams, concentrated loads and stability; the spherical expression demonstrates only that $m_E \propto r^2$ cannot be assumed indefinitely.

The USGS estimates 2025 global helium production at 190 million cubic meters. Its conversion convention, 4.7 kg per 27.737 m^3 at reference conditions, gives approximately 32.20 $kt\,yr^{-1}$. A fleet of one thousand central-case platforms, representing 1 GW of IT, would hold 26.51 kt of helium: approximately 82% of that annual global flow. This reference volume must not be confused with the volume of the expanded gas at 22 km.

This initial inventory is not annual consumption. The latter depends on losses, new commissioning and recovery:

$$\dot{M}_{\text{supply}} = \lambda M_{\text{fleet}} + \dot{M}_{\text{new fills}} - \dot{M}_{\text{recovered}}.$$

The comparison reveals an industrial scaling constraint, without allocating all global production to this use alone. Hydrogen is a physical alternative: at identical pressure and temperature, the gain in net lift is approximately 7.96%. It does not represent an equivalent reduction in total cost. Materials, equipment separation, leakage, ignition prevention and certification must be reassessed.



F15. Modeled mass breakdown and initial helium inventory, compared with the 2025 global production flow estimated by USGS 2026 (190 million m³ at the reference conditions stated in that document). Initial fill inventory is not annual consumption. Hydrogen and gas recovery alter this constraint.

Economics and demonstrations required before 2030

The model's exploratory coefficients yield a central investment of 35.67 million USD per MW of IT, including 18 million USD of IT equipment. This is neither a quotation nor an industrialized production cost. IT is already included in this total; adding it again would constitute double counting. The economic scenario then uses 2 million USD of annual expenditure and hypothetical availability of 0.96, with IT replacement after three or six years depending on the case.

Electrical sizing covers full load. Average billable utilization, set elsewhere to $u = 0.75$, reduces the economic denominator and does not automatically allow a reduction in the power reserved for computing

jobs. With availability a, efficiency η and duration Δt, useful operations are $F = \eta P_{\mathrm{IT}} u a \Delta t$. These conventions make installed infrastructure and the service sold comparable.

Daytime-only IT operation can reduce storage, but changes the product. For a productive fraction f, annual computation becomes $F_{\mathrm{day}} = f \eta P_{\mathrm{IT}} u a \Delta t$; fixed costs per operation increase approximately as $1/f$ before resizing. Maintaining annual volume requires more peak power. Propulsion, telecommunications and safety functions remain necessary at night: shutting down the accelerators does not eliminate all batteries.

A credible pathway begins by measuring the dominant coefficients: actual envelope mass under load, seasonal energy balance, equipped-vehicle drag, heat transfer, gas leakage and maintenance time. It then proceeds to an integrated demonstration capable of delivering a measured contractual service. SACP may retain value for deferrable workloads, exceptionally constrained locations or services combining computing and aerial infrastructure. Its general competitiveness in 2029–2030 remains a conditional hypothesis, whose validation requires experimental closure of both the system and its financing.

The economic model allocates 2 M USD/MW/year to operations and assumes replacement of IT alone when replacement is selected. Recovery, intervention and return to altitude must be covered by this allocation or added explicitly; the absence of quotations for them makes this a favorable case.

Orbital computing in 2029–2030: a measurable industrial frontier

A favorable architecture built from the computing requirement

Low-Earth-orbit computing has a conditional window of competitiveness in the model, but this window depends simultaneously on mass per unit of IT capacity, the full launch price, the cost of accessible terrestrial sites and accelerator replacement. A conclusion based solely on solar irradiance or price per kilogram launched would be incomplete. The relevant objective for 2030 is to determine where this window opens, how much capacity can actually pass through it, and which workloads retain their efficiency when distributed in space.

The favorable architecture selected uses orbits near the dawn–dusk plane, deployable photovoltaic arrays, a short thermal loop, high-emissivity radiators and an optical connection to a relay network. Tightly coupled workloads remain within one satellite or a compact formation. Long-distance communications carry datasets, checkpoints and results, rather than all internal accelerator transactions. This distinction avoids conflating the constraints of a supercomputer with those of a satellite communications service.

The unit of analysis is one MW of nominal IT power; utilization, availability and software efficiency then determine useful FLOP. This convention must be maintained through the final table: photovoltaic MW, peak IT MW, average IT MW and MW of useful computing are not interchangeable. A large fleet can execute independent requests without forming a single synchronous cluster. A constellation's total power therefore does not, by itself, establish its ability to train a single model.

For independent inference and batch processing, task decoupling provides useful tolerance to local interruptions. For tightly synchronized training, inter-accelerator bandwidth and collective stalls are decisive. A separation of 100 m introduces only 0.33 microseconds of propagation, compared with 16.7 ms over 5,000 km. Transceivers, switching, serialization and protocols must still be added; short distance can resolve the geometric component without automatically resolving the entire network.

What industrial announcements actually establish

In its June 2026 SEC filing, SpaceX presents a Starship V3 capacity of 100 t and a future generation of 200 t. The document specifies that these are theoretical performances calculated under certain conditions, and anticipates the start of AI satellite deployment in 2028. It reports booster reuse, but still identifies ship recovery and reflight as a milestone yet to be achieved at that date. Promised performance defines a design scenario; it cannot be equated with an available operational launch rate. SpaceX, SEC filing.

The AI1 page consulted announces up to 250 kW peak, 175 kW average and vehicle power efficiency of 75 kW per tonne. The last figure does not explicitly state its time basis. If it uses average power, the implied mass is 2.33 t; if it uses peak power, it is 3.33 t. Relative to the 175 kW average, these interpretations give 13.33 and 19.05 kg/kW. This calculation bounds a documentary ambiguity; it does not replace an official mass specification. The announced manufacture of thousands of units as early as late 2027 remains an industrial target. SpaceX, AI1.

Suncatcher contributes original work on close formation, optics, radiation and transportation economics. Its preprint notably models 81 satellites within a radius of 1 km at an altitude of 650 km. Its cost study examines a trajectory toward less than 200 USD/kg in the mid-2030s; it compares amortized launch cost with terrestrial electricity, rather than providing a complete LCOC. Attributing guaranteed overall parity by 2030 to this work would therefore exceed the published result. Google, Suncatcher preprint.

Another approach deserves separate consideration: integrating computing, solar generation and radiation surfaces into the same panels. An April 2026 preprint targets approximately 112 kW per tonne at the system level, with a 150 t concept and a structure extending several kilometers. Shared structures can reduce mass; this approach still requires thermal, structural and radiation qualification. It justifies a sensitivity near 10 kg/kW, without justifying presentation of that level as demonstrated performance. Gaalema, Indyk and Staley, 2026.

Closing the thermal and mass balances

The common model sizes all surfaces per nominal kW of IT. With a total electrical-power-to-IT-power ratio $b = 1.15$, photovoltaic efficiency $\eta_{PV} = 0.30$, degradation and loss factor $d = 0.90$, illumination fraction $f_s = 0.98$ and irradiance $G_0 = 1361$ W/m²:

$$a_{PV} = \frac{1000b}{G_0 \eta_{PV} d f_s} = 3.1934 \text{ m}^2/\text{kW}_{IT}.$$

This expression represents a favorable assumption of nearly continuous operation. A Sun-synchronous orbit is not automatically eclipse-free. For a circular orbit at 500 km, the approximate condition for avoiding occultation by a cylindrical Earth shadow is

$$|\beta| > \arcsin\left(\frac{R_E}{R_E + h}\right) \simeq 68.0^\circ.$$

This condition must be verified throughout the year and using the actual geometry. The central model adds no battery for a long eclipse; this simplification must be removed if the selected trajectory requires one. A reserve for transients and nighttime autonomy require different sizing. If an eclipse of duration t_e requires uninterrupted computing, the minimum storage mass becomes

$$m_b = \frac{bP_{IT}t_e}{e_b \eta_b D},$$

where e_b is system specific energy, η_b is discharge efficiency and D is the usable energy fraction. The solar generator must also recharge this reserve, including losses; adding only a battery does not close the energy balance.

In vacuum, heat leaves the radiator by radiation. Chips do not radiate their power directly into an ideal reservoir: interfaces, heat exchangers, pipes and heat-flux distribution impose temperature differences. Defining q''_{abs} as absorbed environmental input, the net flux is

$$q''_{net} = \varepsilon\sigma T_r^4 - q''_{abs}, \qquad a_r = \frac{1000b}{q''_{net}}.$$

For $\varepsilon = 0.90$, $T_r = 350$ K and $q''_{abs} = 100$ W/m², the result is $q''_{net} = 665.82$ W/m² and $a_r = 1.7272$ m²/kW IT. The 100 W/m² input is a scenario parameter; terrestrial view factors, albedo, orientation and occultations must replace it in a mission study. Assuming only a 3 K background temperature would remove part of the physical problem.

The area convention is that of an emitting surface. If both sides participate, their areas and thermal environments must be counted separately; areal mass must remain defined under the same convention. The central mass balance is then

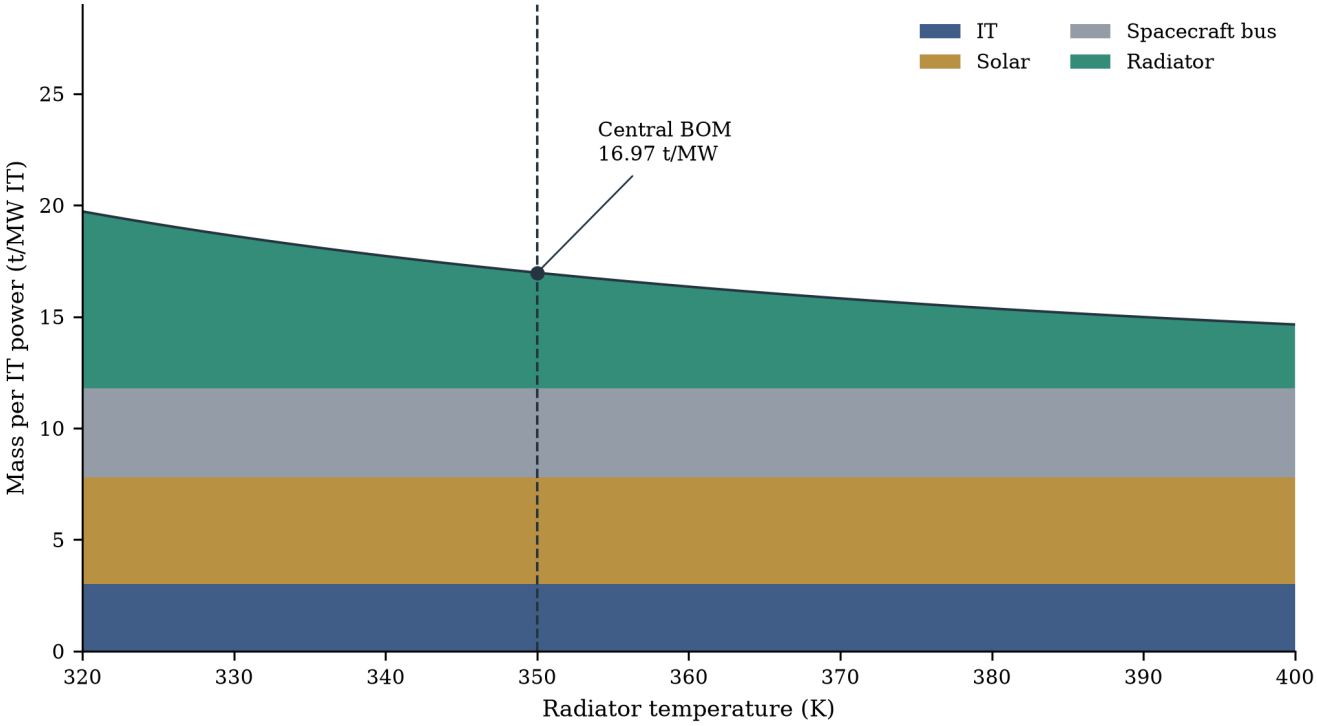
$$m_* = m_{IT} + m_{bus} + \Sigma_{PV} a_{PV} + \Sigma_r a_r.$$

With $m_{IT} = 3$, $m_{bus} = 4$ kg/kW, $\Sigma_{PV} = 1.5$ and $\Sigma_r = 3$ kg/m²:

$$m_* = 3 + 4 + 4.7901 + 5.1816 = 16.9716 \text{ kg/kW}_{IT}.$$

The spacecraft bus includes control, communications, structure not allocated to the panels, and propulsion. Its value remains an engineering assumption to be verified through a detailed bill of materials, not a qualification result. Array-specific power reported by NASA for existing missions provides useful references, but does not impose a physical ceiling on future large integrated structures. NASA, state of the art in power subsystems.

F12. Mass reduction through higher radiator temperature



Thermal extrapolation at constant IT performance; junction compatibility remains to be qualified.

Mass per MW IT, obtained without changing orbital_design. One tonne/MW equals one kg/kW. The radiator uses one emitting face, ε=0.90, absorbed input 100 W/m² and 3 kg/m²; areas are calculated by the common model. IT 3 kg/kW, spacecraft bus 4 kg/kW, solar 1.5 kg/m². Extension to 400 K is a conditional calculation at constant IT efficiency: it does not demonstrate chip, coolant or material compatibility. Degradation in FLOP/J or a heat pump would alter the optimum.

Figure F12 reduces total mass from 19.72 to 14.65 t/MW as the radiator rises from 320 to 400 K. This extrapolation artificially holds IT FLOP/J constant to isolate the thermal design variable. The actual optimum must incorporate junction temperature, electrical losses, reliability and, if necessary, heat-pump energy. A hotter radiator is an improvement only if the useful computing gained per kilogram compensates for these penalties.

Insertion energy and transportation cost: two distinct metrics

Starting from an object at rest on the surface of a spherical, nonrotating Earth, the minimum mechanical energy per kilogram required to reach a circular orbit of radius $r = R_E + h$ is

$$\Delta e = \frac{\mu}{R_E} - \frac{\mu}{2r}.$$

At 500 km, this is approximately 33.56 MJ/kg. Raising one kilogram to 22 km without orbital velocity ideally requires only

$$\Delta e_{22} = \mu\left(\frac{1}{R_E} - \frac{1}{R_E + 22\,000}\right) = 0.2153 \text{ MJ/kg}.$$

The ratio is close to 156. These values are mechanical energy differences, not fuel consumption or a complete comparison of services. Actual chemical energy includes propulsion losses, ascent, gravity, drag, stage masses and recovery. A Hohmann-transfer formula between two orbits does not describe departure from rest at the surface.

Circular velocity at 500 km is $v_c = \sqrt{\mu/r} \simeq 7.62$ km/s, but it is likewise not a launch vehicle's total velocity budget. Delivered price per kilogram depends primarily on reuse rate, capital tied up, maintenance, mission load factors and losses. A substantial economic reduction therefore remains compatible with high mechanical energy. Conversely, low stratospheric lifting energy does not prove the competitiveness of an aircraft maintained on station for several years.

Central LCOC and its competitiveness thresholds

The shared model evaluates six years of operation from commissioning, with a 10% discount rate. It uses 18 M USD of IT equipment per MW, initial efficiency $\eta_0 = 5 \times 10^{11}$ FLOP/J, utilization of 0.75 and idle power consumption at 20% of nominal load. Annual efficiency improvement available in new generations is set at 15%, but is realized only when chips are replaced. The existing satellite does not spontaneously benefit from technological progress.

For a replacement interval k, the generation installed in year t is $v_t = k\lfloor (t-1)/k \rfloor$. Discounted FLOP, for one nominal MW, are

$$F_{PV} = \eta_0\, 10^6 u\, a\, \chi\, Y \sum_{t=1}^{6} \frac{(1+g_\eta)^{v_t}}{(1+r)^t},$$

where Y is the annual duration in seconds, a is availability and χ is relative software efficiency. The central model uses $a_g = 0.995$ on the ground, $a_o = 0.97$ in orbit and $\chi = 1$. An additional network or radiation penalty must be introduced into χ when observed; availability must not count the same loss twice.

For orbit, central capital expenditure excluding launch reaches 26.6074 M USD/MW. It includes IT, panels, radiators, spacecraft bus, qualification, ground communications and integration. Transportation adds $m_* p_L/1000$ M USD/MW when p_L is expressed in USD/kg. At 200 USD/kg:

$$C_{orb} = 26.6074 + 16.9716 \times \frac{200}{1000} = 30.0017 \text{ M USD/MW}.$$

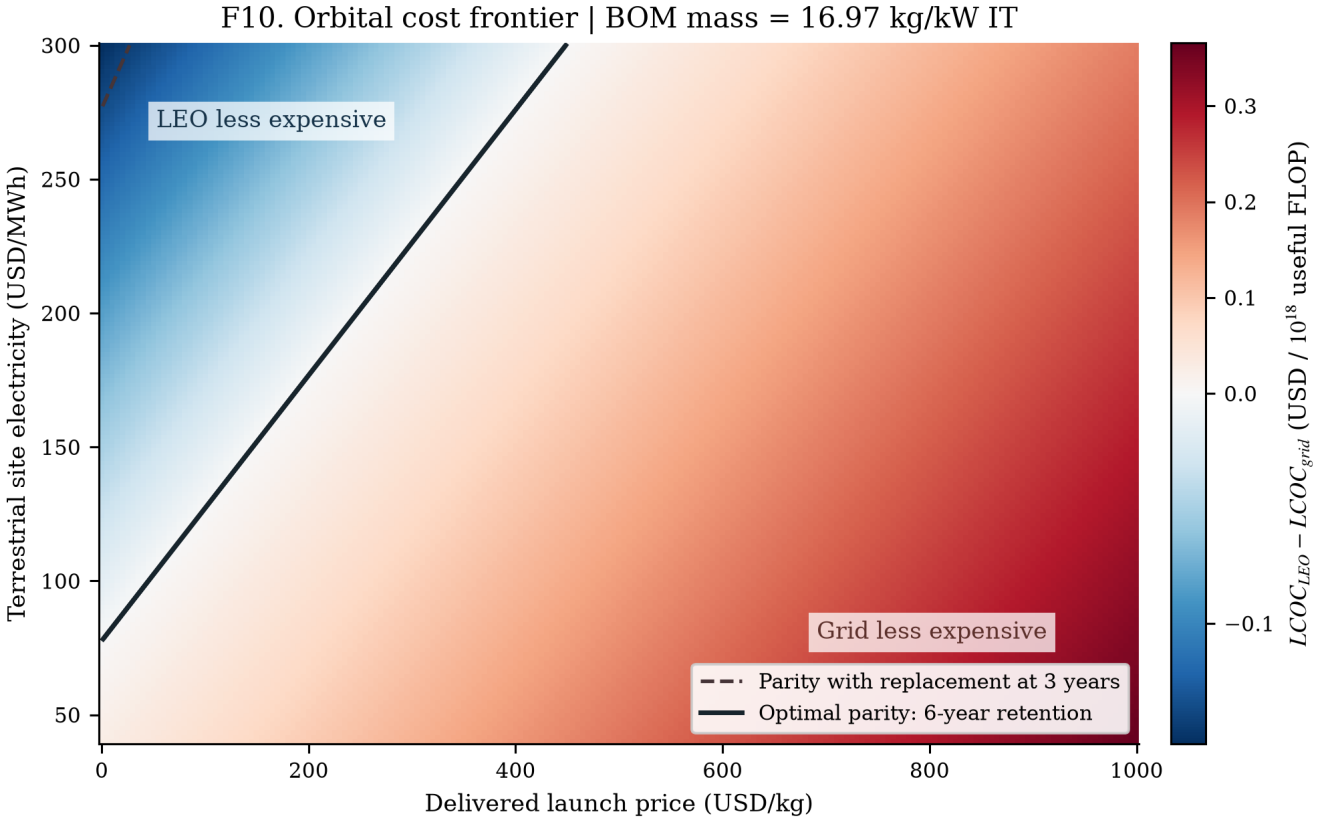
A nominal 100 MW fleet therefore represents approximately 1,697 t, 339.4 M USD of transportation and 3.000 billion USD of initial capital. These are results of the generic scenario; they are neither a SpaceX quotation nor an official estimate for AI1. Orbital operations add 1.1 M USD/MW/year. Orbital

replacement here replaces the complete vehicle; the terrestrial site retains its infrastructure and replaces only IT.

The grid-connected site uses 8 M USD/MW of infrastructure, a PUE of 1.15 and 0.8 M USD/MW/year of fixed operating expenditure. Its annual energy accounts for idle consumption:

$$E_g = 8760\,PUE\,[u + (1-u)f_{idle}] = 8059.2 \text{ MWh/MW/year.}$$

LCOC is the ratio of all discounted expenditure to F_{PV}. At 80 USD/MWh on the ground and 200 USD/kg in orbit, the minima between six-year retention and three-year replacement are **0.6301 and 0.6964 USD per 10^{18} useful FLOP**, respectively. Both architectures then select retention. The orbital premium is approximately 10.5% in this configuration; it is not a universal property of space.



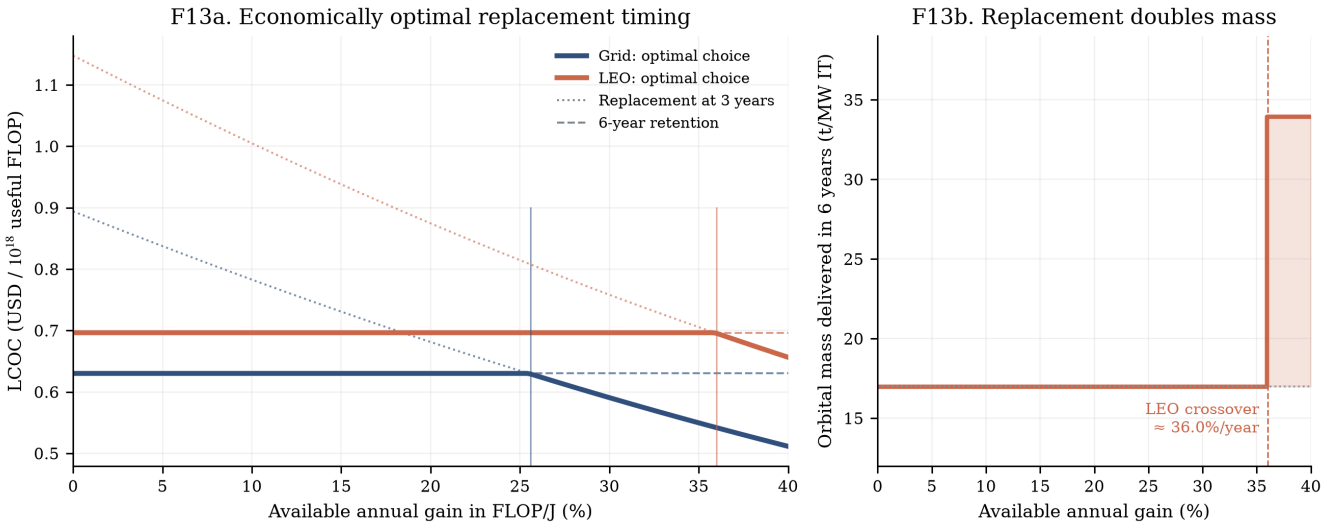
6-year horizon; hardware improvement 15%/year, realized only at replacement.

Conditional comparison from model.py: LEO cost minus grid cost. The bill of materials at 350 K gives 16.972 kg/kW IT. Each architecture chooses between replacement after three years and six-year retention. At 15%/year, both retain their hardware for six years here. The dashed threshold imposes three years on both architectures. Component and launch prices are assumptions, not SpaceX quotations. Availability and performance follow the common model exclusively.

For a fixed replacement strategy, orbital LCOC is affine in p_L and terrestrial LCOC is affine in electricity price. Figure F10 calculates their difference, then selects the best strategy. The break-even launch price is 4.87 USD/kg against a site at 80 USD/MWh, 146.00 USD/kg at 150 USD/MWh, 246.81 USD/kg at 200 USD/MWh and 448.43 USD/kg at 300 USD/MWh. At 200 USD/kg, orbit becomes cheaper when electricity at the comparison site exceeds approximately 176.78 USD/MWh, with all other assumptions unchanged.

A negative threshold means that this orbital configuration remains more expensive even without a launch bill; reducing mass is then insufficient. Manufacturing, qualification or operating costs must fall, availability must improve, or a different workload and comparison site must be selected. Conversely, sites genuinely constrained by expensive energy may open an orbital window before launch reaches its theoretical marginal cost. Variants with a four-year life, a less expensive platform or different masses yield other thresholds: they must be presented as sensitivities, never mixed with the central results.

IT replacement and physical obligations



Grid: 80 USD/MWh; launch: 200 USD/kg. Retained hardware does not improve spontaneously.

Choice between replacement after three years and six-year retention over a six-year horizon. Annual hardware improvement is realized only upon replacement. The thick curve is the minimum of the two strategies. The grid-connected site replaces only IT; LEO replaces the complete vehicle, consistent with the shared model. Delivered mass includes the initial satellite and, where applicable, its replacement; it excludes launch losses and unscheduled replacements. Deorbited mass follows at end of life. No modular robotic servicing is assumed.

Figure F13 varies the available annual improvement in FLOP/J from 0 to 40%. Three-year replacement becomes optimal at approximately 25.6%/year for the terrestrial site and 36.0%/year for orbit, on the grid studied. The difference results from replacing the entire vehicle in orbit. At 15%/year, the expenditure on a second satellite is not offset, over this horizon, by its additional discounted useful computation.

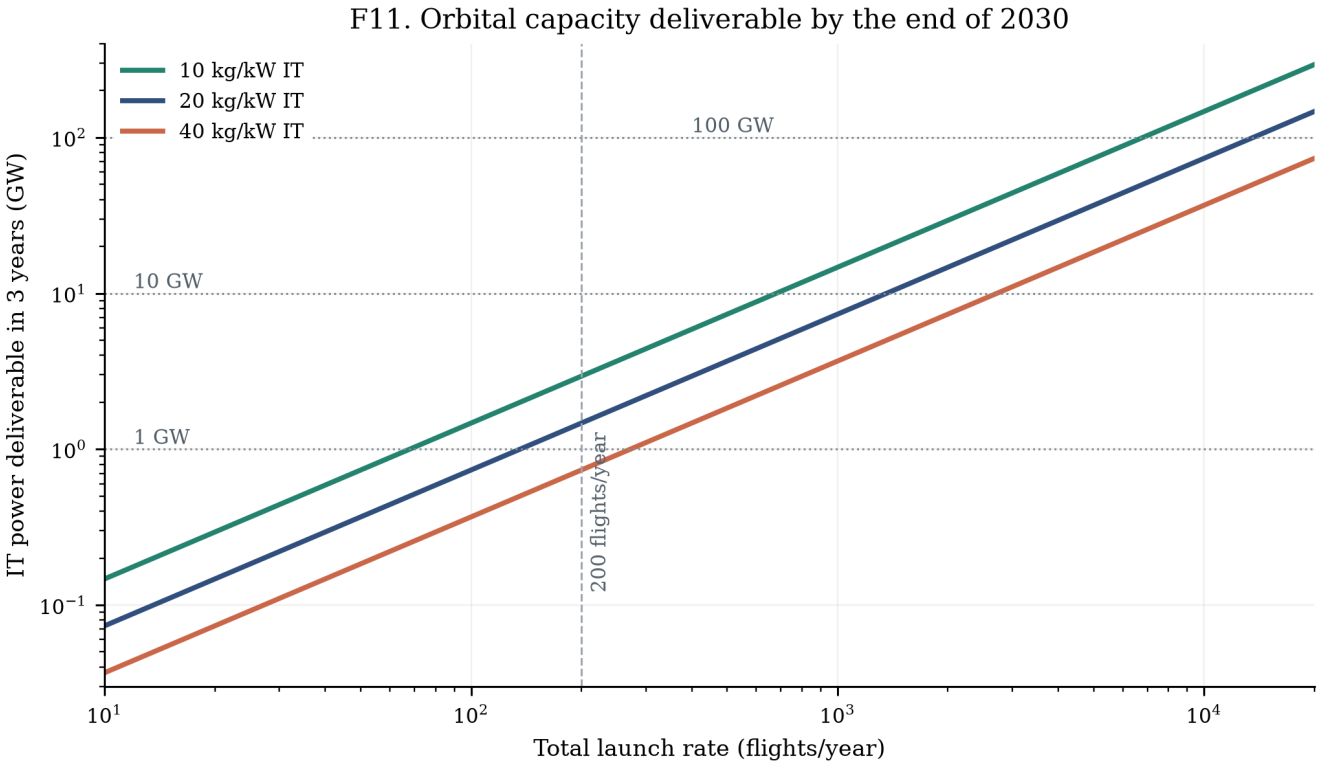
The decision is also physical: replacement after three years doubles the mass delivered over six years, from 16.97 to 33.94 t/MW, excluding losses and unplanned replacements. A modular architecture could retain panels and radiators while changing IT, but must then demonstrate assembly, interfaces, heat transfer and orbital servicing. Assigning it the favorable cost of a simple board replacement while retaining a monolithic architecture would be inconsistent. A long structural life does not guarantee a long economic life for accelerators.

Launch rate distinguishes a demonstrator from grid substitution

For an effectively deliverable payload M_L, a fraction f_{AI} of missions allocated to computing and delivery probability p_s, the annual increase in nominal power is

$$\Delta P_{IT} = \frac{N_{flight} M_L f_{AI} p_s}{m_*}.$$

Figure F11 assumes 100 t per mission, a 50% AI allocation, a 98% delivery success rate and three years of construction, from 2028 to 2030. Masses of 10, 20 and 40 kg/kW are sensitivities distinct from the central bill of materials at 16.97. Stowed volume, adapters, inclination and altitude may constrain M_L below its announced nominal value.



100 t/flight; AI allocation 50%; delivery 98%; construction 2028–2030, without scheduled retirement.

Nominal power delivered by N × 100,000 kg × 0.5 × 0.98 × 3 years / specific mass. Masses of 10, 20 and 40 kg/kW are scenarios, distinct from the central bill of materials. Non-AI missions are included in the total launch rate. No in-orbit failure or scheduled retirement occurs during the first three years; a longer build-out must track cohorts. The payload actually deliverable to the selected orbit and stowed volume must be verified. The 200 flights/year line is a visual reference, not a forecast.

At 20 kg/kW, building 100 MW in three years requires approximately 14 launches per year in total; one GW requires 136, ten GW 1,361 and one hundred GW 13,605. A branch at 10 kg/kW and 200 t per flight divides these rates by four. The difference between 100 MW and 100 GW is therefore not a simple commercial extrapolation: it changes the required number of vehicles, factories, launch slots and operational systems.

FAA documents analyze up to 25 annual launches at Boca Chica and 44 at LC-39A. These envelopes may evolve and other sites may contribute, but they make visible the infrastructure work required before thousands of annual missions can be reached. They are neither guaranteed launch rates nor a permanent ceiling for 2030. FAA, Boca Chica, FAA, Kennedy study.

A new fleet does not undergo scheduled replacement in its first year. Correct tracking uses cohorts, $P(t) = \sum_y \Delta P_y S(t-y)\mathbf{1}_{t-y<L}$, with survival S and service life L. Only at maturity does the minimum replacement requirement approach P/L. Price per kilogram and launch rate must remain consistent: a price assumption based on extensive reuse requires a market and an industrial system capable of supporting that activity.

Decision criteria before widespread deployment

Radiation and reliability must be measured at the level of validated computation. For a group of n units whose independent failures each interrupt the entire task, the mean time between interruptions becomes approximately $1/(n\lambda)$. Local restart, replication and checkpoints can substantially improve this result, but consume memory, bandwidth and energy. Useful efficiency must incorporate these costs, rather than treating a tolerated total dose as a guarantee of application availability.

End of life also closes the industrial balance. One hundred GW at 20 kg/kW, replaced every four years, would imply 500,000 t/year of replaced vehicles. This calculated flow is not a forecast of mass converted entirely into aerosols; it shows why the reentry plan must be studied at fleet scale. NOAA has measured metals from reentries in stratospheric aerosols, while the consequences of a substantial increase remain to be quantified. NOAA and associated PNAS publication.

A credible orbital strategy for 2030 must therefore satisfy four conditions jointly: lower LCOC for the selected workload, a detailed mass budget consistent with surfaces and temperatures, an actually deliverable launch rate, and demonstrated computing availability. The model identifies conditions under which orbit wins, particularly against expensive terrestrial electricity sites and with sufficiently economical platform manufacturing. It also identifies the variables that reverse the ranking. An initial 100 MW system may be rational before global substitution at tens of GW becomes feasible; the investment decision must concern this specific tranche and its customers.

Compute Cost, Deployment Schedule, and Financing

Economic performance must be measured against a defined quantity of useful compute, delivered on a specified date and in accordance with its contract. A low energy price guarantees neither high accelerator utilization nor debt repayment capacity. The parameters below are reproducible model assumptions, not supplier quotations or financing offers.

1. LCOC on a Consistent Basis

For six operating years from commissioning, the levelized cost of compute is

$$\mathrm{LCOC} = \frac{I_0 + \sum_{t=1}^{6} (O_t + I_{R,t})/(1+r)^t - S_6/(1+r)^6}{\sum_{t=1}^{6} C_t/(1+r)^t}.$$

I_0 denotes initial investment, O_t operating expenditure, $I_{R,t}$ replacements, and S_6 residual value, assumed here to be zero. Discounting the denominator does not imply that physical FLOP disappear over time: this convention calculates the constant price that equates discounted revenue and costs. With $R_t = pC_t$, solving $\mathrm{NPV} = 0$ gives $p = \mathrm{LCOC}$, before taxation and contractual differences.

Assumption per MW IT	Scenario value
Initial IT equipment	USD 18 million
Terrestrial infrastructure	USD 8 million
Electricity price	USD 80/MWh
Annual fixed expenditure	USD 0.8 million
PUE	1.15
Compute utilization u	0.75
Idle power fraction β	0.20
Availability a	0.995
Initial efficiency under load η_0	5×10^{11} FLOP/J
Discount rate r	10%
Horizon	6 years

Electricity consumption is modeled as

$$P_{\mathrm{IT,\,avg}} = 1\,\mathrm{MW}\,[u + (1-u)\beta] = 0.8\,\mathrm{MW},$$

$$E_{\mathrm{fac}} = 0.8 \times 1.15 \times 8760 = 8059.2\,\mathrm{MWh/year}.$$

Electricity therefore costs USD 644,736/year; total OPEX reaches USD 1.444736 million/year. Under a conservative convention, energy costs are not reduced by downtime, whereas useful compute is. Initial annual output is

$$C_1 = \eta_0\, 10^6\, u\, a\, 31536000 = 1.176687 \times 10^{25}\,\mathrm{FLOP}.$$

Two policies are compared: retention for six years or replacement at the end of year three, at a cost of USD 18 million. The assumed technological progress of 15% per year improves installed equipment only when it is replaced. The performance multiplier is

$$w_t = (1+0.15)^{k\lfloor (t-1)/k \rfloor}, \qquad k \in \{3, 6\}, \qquad C_t = C_1 w_t.$$

The model yields USD 0.6301 and 0.7307 per 10^{18} useful FLOP for retention and replacement, respectively. Retention is therefore selected in this particular case. These results assume that older hardware continues to serve the selected workload without a price decline or loss of compatibility; this commercial assumption must be tested separately.

A GPU-hour price is comparable only after conversion:

$$p_{18} = \frac{p_{\mathrm{GPUh}}\, 10^{18}}{3600\, F_{\max}\, u\, \xi}.$$

ξ represents execution losses at fixed precision, network, and task. A reserved machine may incur rental charges during unused hours. A peak benchmark therefore does not automatically establish the cost of a completed task, and LCOC is not a forecast of the selling price.

2. Delivery Date Has Its Own Economic Value

The LCOC of equipment commissioned in 2029 covers six operating years here, through the end of 2034 if operations begin at the start of 2029. It does not represent the MWh consumed during 2030 alone. Comparing this economic indicator with an annual national forecast requires keeping their respective horizons distinct.

To measure the contribution before the end of 2030, the appropriate quantity is instead

$$C_{2027:2030} = \int_{2027}^{2031} K_{\mathrm{IT}}(t)\,\eta(t)\,u(t)\,a(t)\,\xi(t)\,dt.$$

In a deliberately constant example, with neither replacement nor ramp-up, 1 MW available at the beginning of 2027 delivers 4.706748×10^{25} FLOP before the end of 2030, compared with 2.353374×10^{25} if it becomes available at the beginning of 2029. This factor of two isolates the schedule effect; it does not assume that equipment purchased in 2029 would actually have the same performance as equipment purchased in 2027.

A delay also increases interest during construction and may postpone cash receipts. A comparison made at the commissioning date removes these differences; it must therefore be supplemented with NPV at a common decision date, including development and construction expenditure, customer advances, and revenue at their actual dates. Purchasing accelerators long before power becomes available also exposes them to economic obsolescence before they generate revenue.

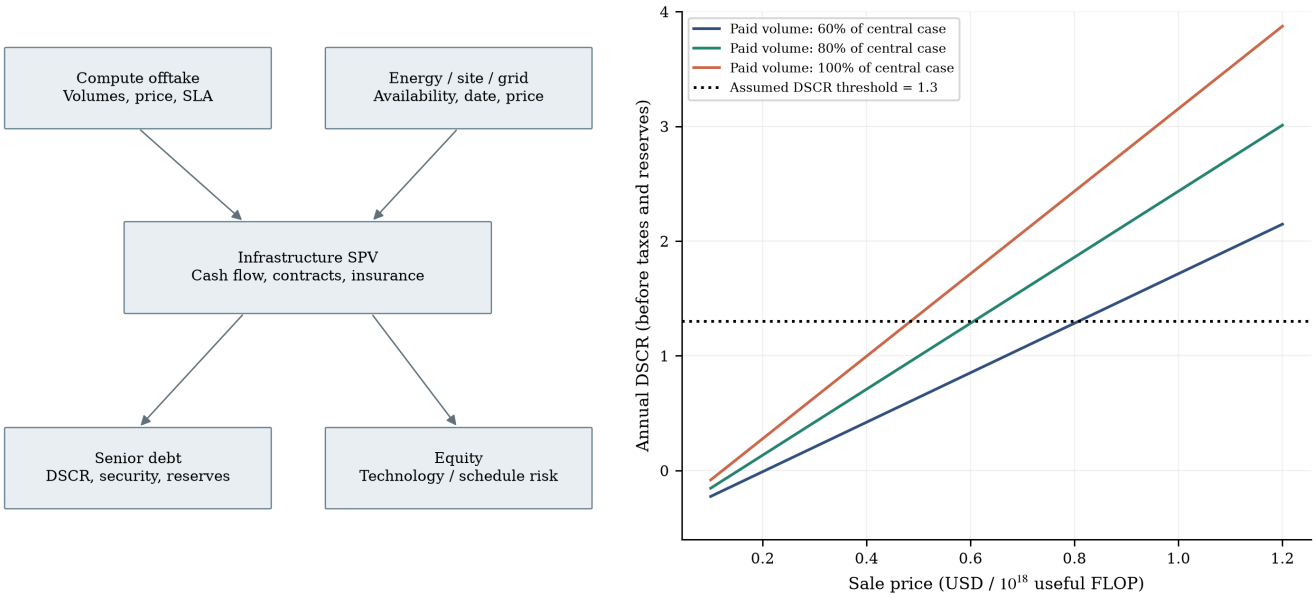
3. From the Compute Contract to the Project Company

The SPV holds the project's assets and contracts. In limited-recourse financing, lenders primarily assess its cash flows, collateral, and contractual allocation of risk; incorporating a company does not eliminate technical risk. World Bank, financing structures.

The compute offtake agreement must specify the task or reserved resource, minimum volume, duration, price, availability, latency, security, acceptance conditions, and consequences of degraded service. Its energy counterpart, the PPA, secures electricity supply or a price hedge according to its terms; it does not, by itself, guarantee continuous data center operation. The schedules of both contracts must align with grid connection, IT deliveries, and the service being sold.

A minimum payment obligation may reduce commercial risk, but its value depends on the customer's creditworthiness and termination rights. A customer that has promised to pay only after a successful demonstration does not implicitly finance the risk that the demonstration fails. Construction guarantees, insurance, reserves, and operating responsibilities complete this analysis; they must not be assumed to be free.

The economic life of components must also be aligned with the duration of commitments. The building, electrical substation, and fiber may remain in service after accelerator replacement; their associated cash flows and collateral therefore differ. Accounting separation between durable infrastructure and evolving equipment improves the interpretation of financing without eliminating operational interdependencies. A long-term contract requires either a precisely defined capacity commitment or an upgrade rule and replacement funding. Promising a constant price for a service with rising requirements, without allowing for this obligation, transfers technical debt into future years. Credit analysis must also incorporate payment delays and service penalties: an issued invoice is not yet cash available to meet a bank payment.



F21. Contract structure and debt-coverage sensitivity. Hypothetical example: CAPEX 26 M USD/MW IT, debt share 60%, interest rate 7%, six-year annuity, and grid-case OPEX. Taxes, working capital, reserves and commercial obsolescence remain to be added to the credit model.

4. Debt Coverage and Limitations of the Structure

The coverage ratio compares cash flow available for debt service with the periodic debt payment. World Bank, DSCR.

$$\mathrm{DSCR}_t = \frac{\mathrm{CFADS}_t}{\mathrm{Interest}_t + \mathrm{Principal}_t}.$$

In F21, initial debt is USD 15.6 million. The annual payment is

$$A_D = D\frac{i}{1-(1+i)^{-n}} = 3.272814\,\text{million USD/year}.$$

The simplified representation uses

$$\mathrm{DSCR}(p,f) = \frac{pfQ-O}{A_D}, \qquad p_{\min} = \frac{dA_D+O}{fQ},$$

where $Q = 11766870$ units of 10^{18} FLOP/year and f represents the fraction paid for. For the illustrative threshold $d = 1.3$, the minimum price is USD 0.4844 at $f = 1$, 0.6054 at $f = 0.8$, and 0.8073 at $f = 0.6$. These thresholds exclude taxes, changes in working capital, reserves, and distributions to shareholders. They are not terms offered by a lender.

The first threshold is below full LCOC: covering senior debt does not necessarily remunerate equity. Annual DSCR also fails to detect every refinancing or replacement difficulty. A credit decision must test cash flows after full expenditure, contract duration, obsolescence, and correlated failures. A demonstrator without an operating history does not acquire proven debt capacity merely because a central-case curve exceeds 1.3.

5. Stage Investment and Purchase Information

The transition from 1 MW to 100 MW and then 1 GW must depend on different forms of evidence. The first stage measures availability, maintenance, and actual costs; the second demonstrates replication, operations, and commercial contracts; the third requires an industrial supply chain and power supply compatible with that scale. Proportional CAPEX scaling demonstrates none of these transitions.

The option to wait has value when future information may change the investment decision. Formally,

$$V_{\text{information}} = \mathbb{E}\Big[\max_d \mathbb{E}[V(d,s) \mid Y]\Big] - \max_d \mathbb{E}[V(d,s)] - C_Y.$$

Y represents pilot results, s the future state, and d the decision. Without defensible probabilities, this expression remains a framework rather than a number: scenarios, decision-switching thresholds, and the cost of delay must be presented. The pilot's purpose is to resolve uncertainty that changes the subsequent investment. Fundraising for this demonstration must be distinguished from infrastructure debt based on an already verified and contracted service.

Broadening the Benchmark: Coastal Architectures and Distant Space Destinations

Terrestrial Architectures Can Also Evolve

The orbital counterfactual must not be limited to an aging campus in a congested metropolitan area purchasing expensive electricity. It must include terrestrial sites that could be built during the same period: brownfield sites with electrical capacity rights, existing power plants, geothermal resources, new transmission lines already committed, and large renewable portfolios with storage and backup supply. This requirement does not predetermine the outcome; it avoids comparing future space infrastructure with past terrestrial infrastructure.

Dedicated generation may be physically connected behind the meter or contractually supplied through a grid. A neighboring power plant does not guarantee the right to supply the site directly. In the scatter-plot model, the assumed price of dedicated energy includes generation, supply continuity, and delivery; its capital is not added a second time to IT capital. A project study must choose between a complete energy bill and a power plant owned by the SPV with explicit CAPEX, fuel, maintenance, and reserves.

The power-balance equation for generation with storage and a grid connection becomes

$$P_{generation,h} + P_{grid,h} + P_{discharge,h} = P_{IT,h} + P_{aux,h} + P_{charge,h} + P_{curtailed,h}.$$

A technology label cannot replace this equality. A turbine requires fuel supply, mechanical capacity, permits, and maintenance; hydropower depends on water and its competing uses; geothermal generation requires a productive reservoir and available equipment. A new reactor must be assessed against its actual construction and qualification schedule. Existing nuclear capacity and an uncommitted future reactor cannot receive the same availability indicator for 2029.

Coastal Sites, Floating Platforms, and Submersion

Seawater provides high volumetric heat-transport capacity, but the site must fund intakes and outfalls, heat exchangers, pumping, corrosion control, biofouling management, and maintenance. For a water loop:

$$\dot{Q} = \dot{m} c_p \Delta T, \qquad P_{pump} = \frac{\Delta p \dot{V}}{\eta_p}.$$

Discharge flow, local temperature rise, and ecological impacts impose constraints distinct from the volume of water consumed. Low WUE does not imply zero withdrawal. In warm waters, the thermal gradient decreases; a coastal location is therefore not uniformly favorable. At greater depth, pressure and intervention requirements also change hull and replacement costs.

Microsoft reports that the server failure rate of the Natick Northern Isles demonstrator was one-eighth that of the terrestrial comparison group. This observation relates to a specific experiment that included a controlled internal environment. It supports investigating sealed modules; it demonstrates neither a globally lower LCOC nor economics that automatically scale to multiple GW. Microsoft Research, Project Natick.

A barge near an electrical substation may facilitate modular manufacturing; an offshore platform may bring cooling and generation closer together. However, generating electricity offshore does not require moving GPUs offshore: the electrical cable to shore must be compared with data transport and maintenance of the offshore installation. The economic criterion concerns these complete alternatives, including vessel availability and interruptions during replacement.

Beyond LEO: Distance Provides an Initial Screening Criterion

Architecture	Potential benefit	Governing constraint	Role in a 2029–2030 decision
Reinforced grid-connected terrestrial site	Existing maintenance and supply chain	Electrical node, schedule, customer	Contractable phases, selected locally
Terrestrial site with dedicated energy	Choice of supply	Firmness, permits, energy equipment	Complete alternative to a constrained campus
Coastal / floating	Heat rejection and modularity	Delivered energy, corrosion, marine access	Site-specific case, not a universal advantage
Sealed subsea	Controlled environment, cooling	Retrieval and IT obsolescence	Specialized deployment to investigate
SACP at 20–25 km	Low lifting energy, vertical proximity	Buoyancy, storage, wind, structure	Demonstration program subject to system closure
Solar LEO	Onboard generation and space network	Mass, radiators, cadence, replacement	Conditional economic domain and phased scaling
MEO, example at 20,000 km	Wider geometric coverage	Propagation, radiation, launch	Requires a distinct advantage absent from the current model
GEO	Prolonged visibility of a region	Distance and heat rejection, transport	Applications tolerating greater latency
Lunar surface	Physical support and potential local resources	Delivery, local night, dust, communications	Research hypothesis beyond the demonstrated industrial case
Earth–Sun L1 environment	Sustained solar access depending on trajectory	Distance, communications, logistics	Poorly suited to interactive terrestrial services

A selected MEO altitude of 20,000 km already gives approximately 133 ms of vertical RTT to an onboard computer. For GEO at 35,786 km, the bound becomes approximately 239 ms. These values do not represent a round trip between two ground users relayed by a satellite, which adds segments. GEO altitude is documented by NASA, TDRS systems.

At the mean lunar distance of 384,400 km, geometric RTT is approximately 2.56 s; at 1.5 million km, near Earth–Sun L1, it approaches ten seconds. These values follow from $2R/c$, before processing. The distances are documented in NASA lunar data and the DSCOVR mission. They are sufficient to exclude these locations for some interactive SLAs, but not for all asynchronous computation.

The lunar poles do not simply place permanent sunlight beside free cooling. Illumination depends on topography; some areas remain illuminated for long periods, whereas others are persistently shadowed. Transporting heat and energy between them remains an infrastructure problem. Site and illumination-cycle constraints are described in NASA work on the lunar south pole. In the absence of a

compute logistics chain at the scale considered, this study does not assign the location a fabricated USD/FLOP price merely to complete a chart.

Integrated Results: Competitive Domains and the Value of the Schedule

The Central Case Is an Anchor, Not a Forecast

Architecture	**Initial CAPEX per MW IT**	**LCOC USD / 10^{18} useful FLOP**	**Selected replacement policy**	**Major additional condition**
Central grid-connected case	USD 26.00 million	0.6301	Retain for 6 years	Assumed electricity price of USD 80/MWh and an available site
Central LEO case	USD 30.00 million	0.6964	Retain for 6 years	USD 200/kg, mass of 16.97 kg/kW, availability of 0.97
Central SACP case	USD 35.67 million	0.8976	Retain for 6 years	Structure and storage assumed feasible, availability of 0.96

The differences in this table result from a transparent but incomplete model. It excludes, among other items, local taxation, residual value, commercial margins, and unquantified external costs. Equipment, energy prices, mass, and OPEX are not all observed quantities for 2030. Four decimal places facilitate verification of model execution, rather than expressing economic certainty to four decimal places.

IT equipment represents most of the shared capital expenditure. This structure explains why even substantial electricity savings may reduce LCOC only modestly. For the terrestrial case with a fixed replacement policy:

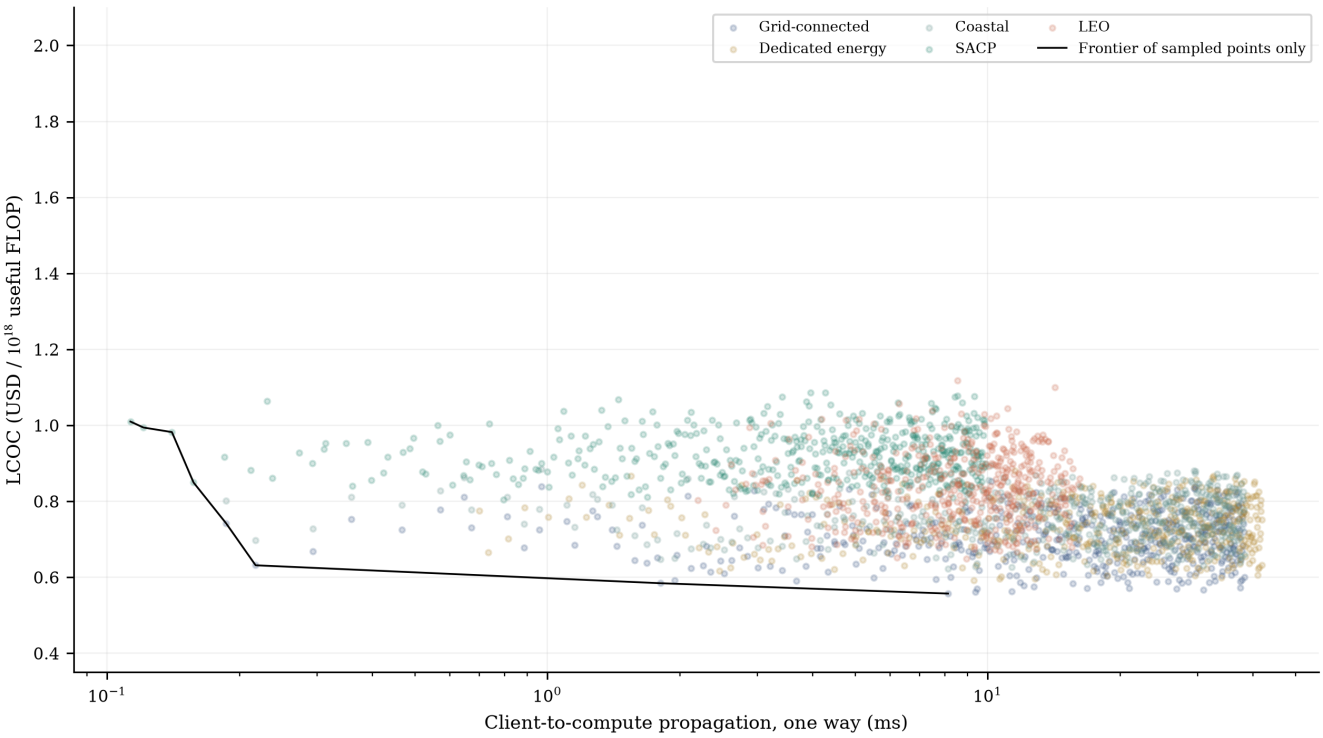
$$\frac{\partial LCOC_g}{\partial p_E} = \frac{E_g \sum_t (1+r)^{-t}}{F_{PV}}.$$

An increase of USD 100/MWh changes cost here by approximately USD 0.0685 per 10^{18} useful FLOP. This sensitivity is specific to the selected efficiency and utilization. If FLOP/J progress sharply reduces required power, capital and interconnects may become even more dominant. Conversely, if firm energy becomes expensive or unavailable by the required date, its marginal value may substantially exceed its average tariff.

Scatter Plots and the Pareto Frontier

A configuration x dominates y when both its cost and latency are lower or equal, with at least one strict inequality. The nondominated set within a domain $\mathcal{D}$ is

$$\mathcal{P} = \{x \in \mathcal{D} : \nexists y \in \mathcal{D},\ C(y) \le C(x),\ L(y) \le L(x),\ (C(y), L(y)) \neq (C(x), L(x))\}.$$



F19. 2,500 exploratory configurations (500 per family, stratified Latin hypercube sampling, seed 2030). Densities and frequencies are neither probabilities nor observations. Local solar costs include assumed firm energy in their price. Telecommunications paths are geometric scenarios. The frontier applies to this sample, without a guarantee of global optimality or feasibility in 2030.

The scatter plot contains 500 configurations per family. Stratified Latin hypercube sampling, with a fixed seed, spans ranges of cost and physical parameters. It does not reproduce a population of data centers and is not a calibrated probabilistic model. Point density depends on parameterization, including the choice to distribute a price or mass linearly; it does not measure a probability of success. The number of points is not a measure of evidential strength.

Family	**Main explored ranges**	**Interpretation**
Grid-connected	Energy USD 40–200/MWh; infrastructure USD 5–15 million/MW; PUE 1.10–1.30	Assumptions for future admissible sites
Dedicated energy	Firm energy USD 60–210/MWh; IT infrastructure USD 6–16 million/MW	Generation already included in the energy bill
Coastal	Energy USD 55–180/MWh; infrastructure USD 8–18 million/MW; PUE 1.08–1.23	Potentially favorable cooling, explicit marine cost premium
SACP	Battery 180–600 Wh/kg; wind 5–20 m/s; envelope 0.3–0.8 kg/m²	Mass and energy closure calculated for each point
LEO	Launch USD 50–700/kg; radiator 330–390 K; variable areal densities and support-system masses	Areas and mass calculated, no implicit qualification

Availability also varies within assumed domains. Client-to-compute routes are parameterized independently, potentially producing favorable combinations that do not yet exist. The plotted frontier is

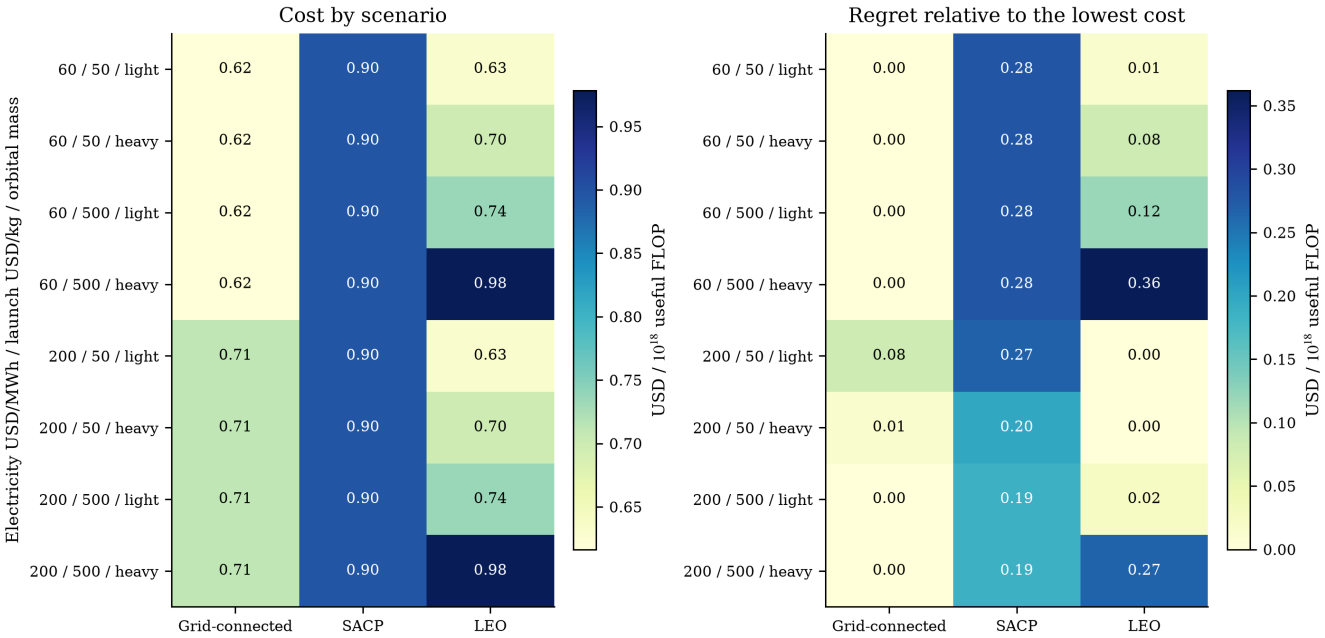
therefore a **conditional exploratory frontier**, not the industry's feasible frontier in 2030. Verification of schedules, sites, and SLAs must eliminate configurations that cannot be built before any investment decision.

This representation nevertheless answers a useful question: which variables must change for an architecture to enter a competitive region? The scatter plot reveals overlaps and influencing variables. Site comparison must then replace independent parameters with integrated offers in which price, schedule, availability, and volume are jointly guaranteed.

Scenario Corners and Regret

Another method defines conflicting scenarios and calculates regret

$$R(a,\omega) = C(a,\omega) - \min_{b} C(b,\omega), \qquad a_{robust} = \arg\min_{a} \max_{\omega} R(a,\omega).$$



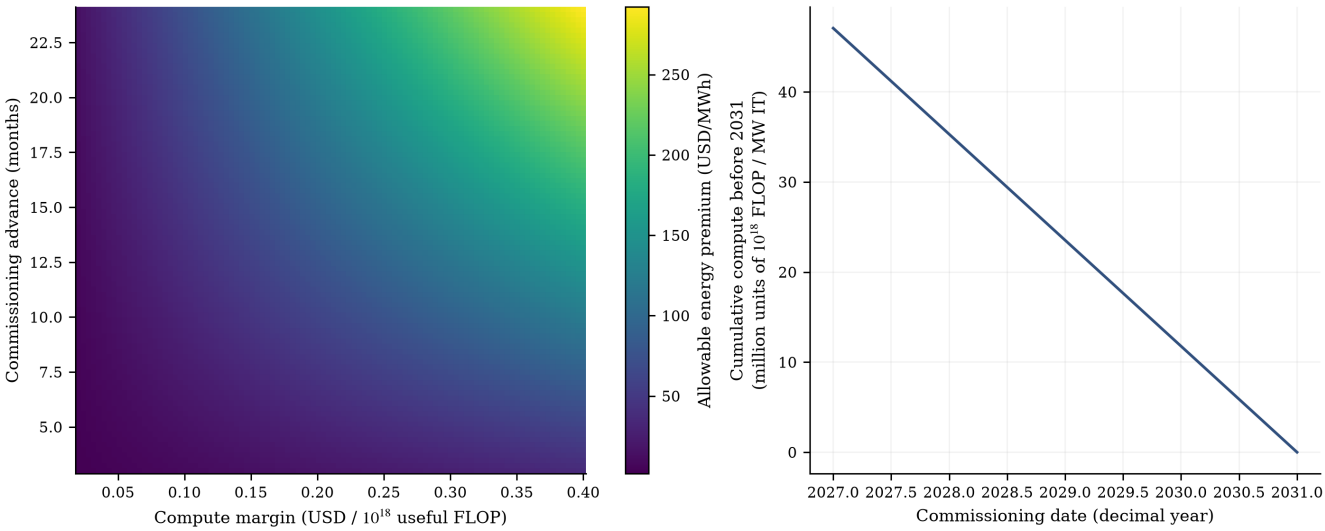
F22. Eight corners of an assumption domain: energy USD 60/200 per MWh, launch USD 50/500 per kg, and light/heavy orbital subsystems. Same IT hardware and service. Conditional economic regret, without probabilities or a 2030 delivery filter. Minimax depends on the selected domain; it does not by itself justify an investment.

F22 uses two electricity prices, USD 60 and 200/MWh, two launch prices, USD 50 and 500/kg, and two orbital variants. The lightweight variant assumes 1 and 1.5 kg/m² for PV and radiators, respectively, plus 3 kg/kW of support systems; the heavy variant assumes 3 and 6 kg/m², plus 8 kg/kW. All other parameters retain their central values. These are neither confidence bounds nor the best and worst possible technologies.

Within this domain, maximum regret reaches approximately 0.0821 for the terrestrial case, 0.2812 for SACP, and USD 0.3618 per 10^{18} FLOP for LEO. The terrestrial case minimizes maximum regret here. However, the orbital case wins in some scenarios with expensive electricity and inexpensive launch. Changing terrestrial infrastructure, orbital manufacturing cost, or deliverable capacity may change this result. The robustness domain must be selected before observing the winner; otherwise, minimax becomes a manufactured ranking.

An operator's rational choice is not necessarily to commit all its capital to the architecture with the lowest regret. It may instead secure terrestrial capacity while purchasing experimental information on an orbital branch. The cost of this option and its ability to open a market must be evaluated separately from the value of an operating fleet.

Capacity Delivered Before 2031 Can Be Worth More Than Low LCOC



F20. Economic cost of delay, 1 MW IT scenario and January 2027–December 2030 window. Left: electricity price premium supported by earlier commissioning at a given unit margin, under an approximation excluding discounting, differential construction cost, and taxation. Right: constant hardware and utilization; a cheaper but later site delivers less compute within the window.

LCOC evaluated over six years after commissioning is insufficient to choose between a site available at the beginning of 2028 and a cheaper site available in 2031. The latter delivers nothing to the customer within the target window. At constant capacity, cumulative compute before a deadline T is

$$F_{before\ T} = \dot{F}_{useful} \max(0, T - t_{service}).$$

The central site produces approximately 11.77 million units of 10^{18} FLOP per year per MW IT. If contribution margin is USD 0.10 per unit, gaining one year represents approximately USD 1.18 million of contribution before other effects. Spread over four operating years and 8,059.2 MWh annually, this supports an electricity price premium of approximately USD 36.5/MWh under the F20 approximation. These figures do not predict the compute price; they make the value of delay explicit for a given margin.

The simplified expression is

$$\Delta p_E^{max} = \frac{m_F \dot{F}_{year} \Delta t}{E_{year} H}.$$

It omits discounting, ramp-up, differential construction expenditure, and changes in demand. It explains why a thermally less favorable site may be the better investment if it is actually available much earlier. The forward-looking benchmark must therefore present two distinct results: operating cost over a common duration and the quantity of service delivered before the strategic deadline.

Decision Thesis for 2029–2030 and a Falsification Program

The Portfolio Depends on the Product Sold

Product	Dominant constraint	Candidates to investigate	Measurement that can change the decision
Tightly coupled training	Fabric, collective availability, checkpointing	Compact campuses with firm power; compact orbital formations if qualified	Useful FLOP per watt in actual distributed training
Regional interactive inference	Total latency, data, peak demand	Sites near customer clusters; nearby platforms if access is reliable	Latency distribution and cost per request at constant quality
Batch inference / independent tasks	Useful-compute price and delivered throughput	Remote sites, flexible energy, favorable orbital branch	Completed work per total cost, including retries
Space-data preprocessing	Avoided downlink volume	Onboard orbital compute	Avoided communications cost and latency
Interruptible compute	Energy windows and value of deferral	Terrestrial solar/wind, daytime SACP requiring resizing	Commercial loss caused by interruption

Computing on data already generated in space has a different counterfactual: downlinking all the data for terrestrial processing may consume bandwidth and time. Such a market can make onboard compute rational without demonstrating that space should host the entire terrestrial cloud. This segmentation creates testable entry points.

For terrestrial sites, the initial prospecting list should target nodes whose documentation indicates a potential change before 2030, followed by requests for firm offers. Northern Finland warrants European investigation; other Nordic regions depend strongly on construction and allocation. US cases must account for pressure on connection queues. Solar-rich regions must be compared on firm price and hourly availability. The atlas provides the facts motivating this due diligence; it guarantees no parcel.

For orbital systems, the experimental priority is to demonstrate total mass, useful computing efficiency, heat rejection, and industrial repeatability simultaneously. A low launch-cost target cannot compensate for an excessively heavy or rapidly obsolete satellite. A 100 MW demonstration must not be presented as a substitute for 100 GW: vehicle production, missions, and replacements differ by several orders of magnitude.

For SACP, a small payload does not validate either the stability of a vehicle hundreds of meters long or its overnight storage. Decisive tests concern actual integrated masses, membrane integrity, station keeping under extreme winds, shutdowns and recovery, and maintenance cost. A daytime workload may reduce IT storage requirements, but the vehicle must still survive and remain controllable at night. This branch must be evaluated as an interruptible service with an appropriately adapted electrical architecture.

Experiments with the Highest Value of Information

Uncertainty	Experiment or due diligence	Decision criterion
Terrestrial MW actually available	Operator offer at the node and connection study	Price, date, power, reserves, penalties, and conditions precedent
Firm energy cost	Multiyear hourly simulation and contract	Energy deficit, backup profile, guarantees
Future IT equipment	Common benchmark across candidates	Same model, precision, and quality; complete electrical measurement
Orbital cooling	Vacuum module with realistic external fluxes	Junction temperature, mass, and pumping power
Orbital vehicle mass	Controlled bill of materials and deployed model	Delivered kg/kW, margins, stowed volume, deployment tests
Launch economics	Firm offer and allocated capacity	Delivered price, number of missions, availability, and compensation
Distributed communications	Training and inference on a representative network	Useful throughput, pauses, error rate, recovery, and tail latency
SACP	Instrumented integrated demonstrator	Closed mass and energy balance under a mission weather profile
Water/carbon acceptability	Site inventory and permits	Hourly and life-cycle impacts, not a national average indicator

The optimal test sequence is economic. Inexpensive information that may invalidate a site should be acquired before extensive engineering. The value of information can be formalized as the improvement in the decision enabled by observation, less its cost. Without a credible probability distribution, invalidation scenarios and spending caps are preferable to an artificially precise expected value.

Environmental Footprint and the Electricity System: Completing the Constraints

Compute cost alone does not represent its social value. Emissions, water, and launch externalities must be expressed against a common functional unit:

$$I_{CO_2} = \frac{M_{manufacturing} + M_{deployment} + \sum_h E_h CI_h + M_{end\ of\ life}}{F_{useful}},$$

$$WUE = \frac{V_{water\ consumed}}{E_{IT}}, \qquad V_{consumed} \neq V_{withdrawn}.$$

Annual contractual electricity attributes, hourly physical emissions, and marginal emissions answer different questions. Combining them into a single number without a defined method would create double counting or an illusion of additionality. Orbital solar avoids operational grid consumption, but manufacturing and transport still require an inventory. Conversely, a country's average carbon intensity does not establish the intensity of the marginal MW actually added to a campus.

This study does not fabricate an environmental score to rank architectures with missing inventories. File F09 assembles qualitative cooling and water constraints. The next quantitative version will require hourly weather, local permits, and life-cycle inventories for the equipment actually selected.

Scientific Conclusion

The strongest proposition for 2029–2030 is a decision framework with falsifiable industrial branches. The scarce resource is useful compute delivered by a specified date and under a specified SLA. Global energy is not a homogeneous reserve; orbit does not provide free electricity; the stratosphere is not a passive refrigerator without structural cost.

The results nevertheless move beyond these general objections. They identify a domain in which LEO becomes cheaper, calculate the cadence required to make it materially relevant, show how IT replacement shifts the frontier, and quantify stratospheric storage constraints. They also provide a method for selecting terrestrial regions without conflating announcements and deliverable power.

No single location is demonstrated to be optimal for all 2030 workloads. Within the central domain studied, moderately priced grid-connected terrestrial infrastructure remains competitive; in domains with expensive energy and favorable orbital manufacturing, orbit wins. Stratospheric architectures require measurable advances in mass, storage, or service design before surpassing these alternatives. The proposed strategy is to build phases with verifiable commitments, preserve options whose thresholds are approaching, and prioritize funding for evidence capable of changing the allocation of the next MW.

Methodological Appendix: Validation, Missing Data and Reproducibility

What Has Been Calculated and What Remains to Be Measured

The figures and economic results are deterministic calculations or parameter explorations. They do not constitute a measurement campaign, service-life validation, structural qualification or a global weather model. The economic model selects between two hardware retention periods; it does not optimize every possible partial-replacement strategy. Availability is an input assumption, not an output from a detailed reliability model. Manufacturing costs may be correlated with mass and industrial production volumes in ways that the linear coefficients do not yet represent.

Fundamental equations require dimensional checks; engineering models also require checks of applicability. A buoyancy solution with a small numerical residual does not demonstrate membrane strength. An average illumination fraction of 98% does not establish battery-free operation. Average electrical availability does not guarantee an IT service-level agreement. These limitations are attached to the relevant results rather than confined to a general disclaimer.

The independent program `verify_models.py` checks cost conversions, mass and energy balances, cyclic storage, the variation of optical power with distance, Pareto calculations and economic thresholds. The figures have been visually reviewed. Verification covers the implementation of the assumptions; it does not validate the future assumptions themselves. The checks led, in particular, to an explicit thermal-plate length, corrected electricity consumption when utilization varies, and replacement of an inadequate storage check with a cyclic balance.

Priority Data for Turning This Study into an Investment

Area	Missing data	Potential effect on the decision
Customer demand	Workloads, paid volumes, prices, SLAs and sovereignty requirements	Changes the optimization metric and addressable market
2030 hardware	Full-system power, density, useful FLOP and contractual cost	Simultaneously shifts energy, mass and capital requirements
Sites	Node-specific offers, firm dates and capacity ramps	Eliminates economically attractive but unavailable regions
Energy	Multiyear hourly profiles, capacity charges and backup	Changes firmness and full cost
SACP atmosphere	Extreme and correlated winds, season and irradiation	Changes propulsion, batteries, trajectory and availability
SACP structure	Allowable stresses, fatigue, ballonets and margins	May invalidate closure at constant envelope areal mass
Orbit	Trajectory, eclipses, external fluxes, radiation dose and errors	Changes batteries, radiators, redundancy and availability
Manufacturing and launch	Cadence, yield, stowed volume and allocated missions	Determines the MW actually deliverable before 2031
Network	Routes, useful throughput, gateway weather and collectives	Reduces useful compute despite high peak FLOP
Finance and environment	Contracts, taxation, reserves and inventories	Changes bankability and external constraints

Regional ranges must not be reused as a database of market tariffs. A subsequent version should replace each range with a documented offer and preserve earlier versions to measure the effect of the change. Probabilistic calibration would require distributions and correlations estimated from appropriate observations. This report therefore assigns no numerical probability to terrestrial, stratospheric or orbital compute winning the comparison.

Structure of the Reproducible Package

The file `model.py` defines the common financial model, `aerostat.py` the stratospheric closure model and `common.py` the figure style. The five scripts `fig_demand.py`, `fig_geography.py`, `fig_orbit.py`, `fig_systems.py` and `fig_decision.py` generate the figures. CSV and JSON files contain the plotted points, assumptions or detailed results. The map baselayer comes from Natural Earth and is distributed with the package to avoid a network dependency during execution.

The ledgers `demand_supply_sources.csv`, `geography_sources.csv`, `orbital_parameter_ledger.csv`, `physical_sources.csv` and `manuscript_references.csv` preserve documentary provenance. The file `parameters_and_results.json` records central results directly from the final functions. Orbital analysis variants are distinguished from the central case. SVG images support vector editing; PNG files provide stable reproduction in the PDF.

To regenerate every figure, install Python 3.10 or later, NumPy and Matplotlib, then run:

```
python -m pip install numpy matplotlib
python code/run_all.py
python code/verify_models.py
```

The scripts do not download data and require no service keys. Monetary parameters remain editable in the functions and CSV files. After an assumption changes, the numerical results in the text must be updated along with the figures, not only the chart containing the modified value.

Standalone Program for the Three Requested Benchmarks

The following program is complete. Copy the entire block into `three_benchmarks.py` and execute it with NumPy and Matplotlib installed. It generates Figures F16, F18 and F19, their vector versions and their data. The remaining figures are generated by the other scripts in the package. The Pareto frontier is the frontier of the parameter sample, not an observed industrial frontier.

```
#!/usr/bin/env python3
"""Three benchmarks, standalone. Requires Python 3.10+, numpy, matplotlib.
Save this entire block as three_benchmarks.py and run python three_benchmarks.py.
Outputs: ./benchmarks_2030/{figures,data}. All forecasts are labeled scenarios.
This script duplicates the exact modular model to make copy/paste possible.
"""
"""Shared figure style and paths. Install numpy and matplotlib to reproduce."""
from pathlib import Path
import json
import numpy as np
import matplotlib
matplotlib.use('Agg')
import matplotlib.pyplot as plt
ROOT = Path.cwd()/'benchmarks_2030'
ROOT.mkdir(parents=True,exist_ok=True)
FIG = ROOT/'figures'
DATA = ROOT/'data'
FIG.mkdir(exist_ok=True)
DATA.mkdir(exist_ok=True)
plt.rcParams.update({'font.family':'serif','font.size':9,
    'mathtext.fontset':'stix','axes.spines.top':False,
    'axes.spines.right':False,'axes.titleweight':'normal',
    'pdf.fonttype':42,'svg.fonttype':'none','axes.labelsize':9,
    'legend.fontsize':8,'xtick.labelsize':8,'ytick.labelsize':8,
    'axes.prop_cycle':plt.cycler(color=['#304f7d','#23836e','#cb6749','#b38832','#86669e'])})
COLORS={'ground':'#304f7d','onsite':'#b38832','coast':'#608f8b',
        'SACP':'#23836e','LEO':'#cb6749','ink':'#1d2834','light':'#e9eef2'}
def save(fig, name, caption=None):
    # SVG supplies vector figures; PNG is used in the manuscript.
    fig.savefig(FIG/(name+'.png'),dpi=280,bbox_inches='tight')
    fig.savefig(FIG/(name+'.svg'),bbox_inches='tight')
    if caption: (FIG/(name+'.txt')).write_text(caption,encoding='utf-8')
    plt.close(fig)
def dump(name,obj):
    def conv(x):
        if isinstance(x,np.ndarray):return x.tolist()
        if isinstance(x,np.generic):return x.item()
        raise TypeError(type(x).__name__)
    (DATA/name).write_text(json.dumps(obj,ensure_ascii=False,indent=2,default=conv))
def grid(ax):ax.grid(alpha=.18,linewidth=.6)
import csv
"""Prospective model; monetary parameters are scenarios, not supplier quotes.
Units: MUSD/MW IT; annual MUSD/MW IT; kg/kW IT; seconds; useful FLOP.
Financial comparisons start at commissioning; readiness is a separate gate.
"""
import numpy as np
YEAR=365*24*3600
HOURS=8760
N=6
RATE=.10
IT_CAPEX=18.0
ETA0=5e11
UTIL=.75
```

```
IDLE=.20
ETA_GROWTH=.15
SIGMA=5.670374419e-8
C=299792458.0
def lifecycle(capex,opex,refresh_cost,availability=.99,duty=1.,chi=1.,
              interval=3,rate=RATE,eta_growth=ETA_GROWTH,
              eta0=ETA0,util=UTIL,years=N):
    """Per-MW initial+annual+renewal, improvement only when hardware replaced.
    Annual cash and delivered compute discounted at year end.
    Refresh at end of year 3 benefits years 4-6; residual value set to zero.
    interval=6 retains original hardware for full modeled life.
    """
    t=np.arange(1,years+1)
    df=(1+rate)**(-t)
    vintages=((t-1)//interval)*interval
    perf=(1+eta_growth)**vintages
    repl=np.zeros(years)
    for year in range(interval,years,interval):repl[year-1]=refresh_cost
    pv_cost=capex+float(np.sum((opex+repl)*df))
    pv_flop=eta0*1e6*util*availability*duty*chi*YEAR*float(np.sum(perf*df))
    return {'lcoc':pv_cost*1e6/pv_flop,'usd_per_EFLOP':pv_cost*1e24/pv_flop,
            'pv_cost_MUSD':pv_cost,'pv_flop':pv_flop,'refresh_interval':interval}
def best_lifecycle(capex,opex,refresh_cost,**kw):
    results=[lifecycle(capex,opex,refresh_cost,interval=k,**kw) for k in [3,6]]
    return min(results,key=lambda v:v['lcoc'])
def grid_cost(price_MWh=80.,infra=8.,fixed_opex=.8,pue=1.15,
              availability=.995,chi=1.,**kw):
    u=kw.get("util",UTIL)
    load=u+(1-u)*IDLE
    energy=price_MWh*pue*load*HOURS/1e6
    v=best_lifecycle(IT_CAPEX+infra,fixed_opex+energy,IT_CAPEX,
                     availability=availability,chi=chi,**kw)
    return dict(v,energy_opex_MUSD=energy,initial_MUSD=IT_CAPEX+infra,
                opex_MUSD=fixed_opex+energy)
def orbital_design(T=350.,eps=.90,q_abs=100.,p_aux=1.15,
                   solar_eff=.30,solar_derate=.90,illum=.98,
                   pv_areal_mass=1.5,rad_areal_mass=3.,it_mass=3.,
                   other_mass=4.,pv_cost_area=250.,rad_cost_area=700.,
                   bus_cost_kg=500.,qualification=1.8,gateway=.8,
                   integration=2.,launch=200.,fixed_opex=1.1,
                   availability=.97,**kw):
    """One emitting face convention; q_abs includes environmental load.
    p_aux is electric load/IT; all non-IT mass includes independent margins.
    No long eclipse battery is charged: favorable near-continuous orbit.
    Costs for PV/radiator are installed subsystem scenario coefficients.
    """
    q=eps*SIGMA*T**4-q_abs
    if q<=0:raise ValueError('Nonpositive heat rejection')
    pv_area=1000*p_aux/(1361*solar_eff*solar_derate*illum)
    rad_area=1000*p_aux/q
    m=it_mass+other_mass+pv_areal_mass*pv_area+rad_areal_mass*rad_area
    pv_cap=pv_area*pv_cost_area/1000
    rad_cap=rad_area*rad_cost_area/1000
    bus_cap=other_mass*bus_cost_kg/1000
    launch_cap=m*launch/1000
    cap=IT_CAPEX+pv_cap+rad_cap+bus_cap+qualification+gateway+integration+launch_cap
    v=best_lifecycle(cap,fixed_opex,cap,availability=availability,**kw)
    return dict(v,mass_kg_kW=m,pv_area_m2_kW=pv_area,
        rad_area_m2_kW=rad_area,qnet_W_m2=q,initial_MUSD=cap,
        nonlaunch_MUSD=cap-launch_cap,launch_MUSD=launch_cap,
        parts_MUSD={'IT':IT_CAPEX,'PV':pv_cap,'Radiator':rad_cap,
                    'Bus':bus_cap,'Qualification':qualification,
                    'Ground links':gateway,'Integration':integration,'Launch':launch_cap})
def pareto(x,y):
    """Exact 2D minimization, equal-x groups handled together."""
    x,y=np.asarray(x),np.asarray(y)
    order=np.lexsort((y,x));mask=np.zeros(len(x),dtype=bool)
    best=np.inf;i=0
    while i<len(order):
        j=i+1
        while j<len(order) and x[order[j]]==x[order[i]]:j+=1
```

```
        group=order[i:j];m=np.min(y[group])
        if m<best:mask[group[y[group]==m]]=True
        best=min(best,m);i=j
    return mask
def haversine(lat1,lon1,lat2,lon2):
    a,b,c,d=np.deg2rad([lat1,lon1,lat2,lon2])
    z=np.sin((c-a)/2)**2+np.cos(a)*np.cos(c)*np.sin((d-b)/2)**2
    return 6371*2*np.arcsin(np.sqrt(np.clip(z,0,1)))
def lhs(n,dim,seed=2030):
    """Stratified exploration, not posterior probabilities."""
    rng=np.random.default_rng(seed)
    return np.column_stack([(rng.permutation(n)+rng.random(n))/n for _ in range(dim)])
def validate():
    assert pareto([1,1,2,3],[2,3,1,2]).tolist()==[True,False,True,False]
    assert lifecycle(10,0,0,interval=6)['lcoc']>0
    assert np.isclose(haversine(0,0,0,90),np.pi*6371/2)
    assert grid_cost(100)['lcoc']>grid_cost(50)['lcoc']
    assert grid_cost(util=.9)['energy_opex_MUSD']>grid_cost(util=.5)['energy_opex_MUSD']
    assert orbital_design(launch=500)['lcoc']>orbital_design(launch=100)['lcoc']
    o=orbital_design()
    assert np.isclose(sum(o['parts_MUSD'].values()),o['initial_MUSD'])
    return {'grid':grid_cost(),'LEO':o}
"""Coupled first-order mass/energy closure, not an airworthiness design."""
import math
def isa(z=22000.):
    H=6356766*z/(6356766+z)
    if H<=20000:
        T=216.65;p=22632.06*math.exp(-9.80665*(H-11000)/(287.05287*T))
    else:
        T=216.65+.001*(H-20000)
        p=5474.889*(T/216.65)**(-9.80665/(.001*287.05287))
    return T,p,p/(287.05287*T),p/(2077.264*T)
def closure(U=10.,eb=250.,z=22000.,night=12.,sun_hours=6.,
            env_areal=.5,pv_areal=1.,rad_areal=5.,Ts=350.,plate_L=1.):
    Ta,p,rho,rhohe=isa(z);Tf=(Ts+Ta)/2
    mu=1.716e-5*(Tf/273.15)**1.5*(273.15+110.4)/(Tf+110.4)
    k=.0241*(Tf/273.15)**1.5*(273.15+194)/(Tf+194)
    pr=mu*1006/k
    Re=p/(287.05287*Tf)*U*plate_L/mu
    hc=k/plate_L*.664*math.sqrt(Re)*pr**(1/3)
    q=hc*(Ts-Ta)+.9*5.670374419e-8*(Ts**4-250**4)
    day=24-night
    def vehicle(D):
        a,b=1.5*D,D/2;e=math.sqrt(1-(b/a)**2)
        V=4/3*math.pi*a*b*b
        S=2*math.pi*b*b*(1+a/(b*e)*math.asin(e))
        Psk=.5*rho*.05*math.pi*b*b*U**3/.7
        P=1.15e6+Psk
        Apv=P*(day+night/.85)/(sun_hours*1300*.3*.9)
        E_nom=P*night/(.95*.8)  # Wh
        mb=E_nom/eb
        Qday=1.15e6+P*night/(.85*day)*(1-.85/.95)
        Qnight=1.15e6+P*(1/.95-1)
        Arad=max(Qday,Qnight)/q
        masses=[10000,env_areal*S,pv_areal*Apv,mb,rad_areal*Arad]
        total=sum(masses);cap=(rho-rhohe)*V
        return {'residual':cap-total,'length_m':3*D,'diameter_m':D,
            'volume_m3':V,'mass_kg':total,'helium_kg':rhohe*V,
            'envelope_m2':S,'pv_m2':Apv,'radiator_m2':Arad,
            'battery_kWh':E_nom/1000,'station_W':Psk,
            'load_W':P,'mass_parts_kg':masses,'qnet_W_m2':q}
    lo,hi=5.,2500.
    if vehicle(hi)['residual']<=0:raise ValueError('No closure within bracket')
    for _ in range(65):
        mid=(lo+hi)/2
        if vehicle(mid)['residual']>0:hi=mid
        else:lo=mid
    return vehicle((lo+hi)/2)
def capex(v,IT=18.,env_USD_kg=100.,pv_USD_m2=200.,battery_USD_kWh=150.,
          radiator_USD_m2=500.,helium_USD_kg=50.,integration_MUSD=4.):
    """Cost coefficients are exploratory hypotheses, not observed prices."""
```

```
    d={'IT':IT,'Envelope':v['mass_parts_kg'][1]*env_USD_kg/1e6,
       'PV':v['pv_m2']*pv_USD_m2/1e6,
       'Battery':v['battery_kWh']*battery_USD_kWh/1e6,
       'Heat exchanger':v['radiator_m2']*radiator_USD_m2/1e6,
       'Helium':v['helium_kg']*helium_USD_kg/1e6,
       'Integration':integration_MUSD}
    return sum(d.values()),d
def thermal():
    density=np.linspace(5,150,220);Ts=350;Ta=218;eps=.9
    qleo=eps*SIGMA*(Ts**4-3**4)
    qs=[2*(Ts-Ta),5*(Ts-Ta),10*(Ts-Ta)]
    f,ax=plt.subplots(1,2,figsize=(10,4),layout='constrained')
    ax[0].plot(density,density*1000/qleo,label='Ideal LEO radiation, 350 K',color=COLORS['LEO'])
    for h,q in zip([2,5,10],qs):ax[0].plot(density,density*1000/q,label=f'SACP convection only: h={h}',ls='--')
    ax[0].plot(density,density*1000/(5*(Ts-Ta)+eps*SIGMA*(Ts**4-250**4)),color=COLORS['SACP'],label='SACP h=5 +
        radiation, 250 K background')
    ax[0].set(xlabel='IT floor-area power density (kW / m²)',ylabel='Heat-exchanger area / IT floor area (m² /
        m²)');grid(ax[0]);ax[0].legend(fontsize=6.8)
    t=np.linspace(300,420,180)
    ax[1].plot(t,1e6/(eps*SIGMA*t**4),label='1 MW; emission only')
    ax[1].plot(t,1e6/(eps*SIGMA*t**4-100),label='1 MW; absorbed load 100 W/m²')
    ax[1].set(xlabel='Surface temperature (K)',ylabel='Radiating area (m² /
        MW)');grid(ax[1]);ax[1].legend(fontsize=7)
    save(f,'F16_thermal_area','F16. Consistent normalization of heat-transfer area to rack power density. Left: h
        is an imposed sensitivity parameter, not a property determined by air at −55 °C alone. Right: temperature,
        emissivity and absorbed load determine the required area in vacuum; each emitting surface is counted
        once.')
    dump('F16_thermal_area.json',dict(density_kW_m2=density,leo_ratio=density*1000/qleo,temperature_K=t,leo_area_m2
        _MW=1e6/(eps*SIGMA*t**4-100)))
def deployment():
    mu=3.986004418e14;Re=6371e3;z=np.linspace(0,500e3,500)
    ep=mu*(1/Re-1/(Re+z));h=500e3;vorb=np.sqrt(mu/(Re+h));dE=mu/Re-mu/(2*(Re+h))-.5*465**2
    hstr=22e3;estr=mu*(1/Re-1/(Re+hstr))
    f,ax=plt.subplots(1,2,figsize=(10,3.7),layout='constrained')
    ax[0].plot(z/1000,ep/1e6,label='Increase in potential energy')
    ax[0].scatter([22,500],[estr/1e6,dE/1e6],c=[COLORS['SACP'],COLORS['LEO']],s=40,zorder=3)
    ax[0].annotate('SACP: 0.215
        MJ/kg',(22,estr/1e6),xytext=(90,7),arrowprops={'arrowstyle':'-','lw':.5},fontsize=8)
    ax[0].annotate(f'500 km orbit: {dE/1e6:.1f}
        MJ/kg',(500,dE/1e6),xytext=(160,28),arrowprops={'arrowstyle':'-','lw':.5},fontsize=8)
    ax[0].set(xlabel='Geometric altitude (km)',ylabel='Minimum mechanical energy (MJ/kg)');grid(ax[0])
    payload=np.geomspace(1,1e5,100)
    ax[1].loglog(payload,payload*estr/3.6e6,label='22 km: gravitational potential',color=COLORS['SACP'])
    ax[1].loglog(payload,payload*dE/3.6e6,label='500 km: circular orbit',color=COLORS['LEO'])
    ax[1].set(xlabel='Deployed mass (kg)',ylabel='Minimum mechanical energy
        (kWh)');grid(ax[1]);ax[1].legend(fontsize=7)
    save(f,'F18_deployment_energy','F18. Mechanical lower bounds, not fuel consumption or cost. Aerostat ascent
        draws on buoyancy and atmospheric work; an equivalent speed of √(2gz) is not a propulsive delta-v. A
        favorable Earth-rotation contribution of 465 m/s is credited in the orbital calculation.')
    dump('F18_deployment_energy.json',dict(stratosphere_Jkg=estr,orbit_Jkg=dE,orbital_ms=vorb,payload_kg=payload))
def sacp_cost(eb=250,U=10,env=.5,availability=.96):
    v=closure(eb=eb,U=U,env_areal=env);cap,_=capex(v)
    return best_lifecycle(cap,2.0,IT_CAPEX,availability=availability),v
def scatter():
    n=500;z=lhs(n,12);rows=[]
    for name in ['Grid-connected','Dedicated energy','Coastal','SACP','LEO']:
        for k,x in enumerate(z):
            if name=='Grid-connected':
                cost=grid_cost(40+160*x[0],infra=5+10*x[1],pue=1.10+.20*x[2],availability=.98+.019*x[3])
                latency=(20+6000*x[4])*1.468*1.3/299.792458
            elif name=='Dedicated energy':
                # Price includes generation/firming; infrastructure excludes generator CAPEX.
                cost=grid_cost(60+150*x[0],infra=6+10*x[1],pue=1.10+.20*x[2],availability=.98+.019*x[3])
                latency=(100+6500*x[4])*1.468*1.3/299.792458
            elif name=='Coastal':
                cost=grid_cost(55+125*x[0],infra=8+10*x[1],pue=1.08+.15*x[2],availability=.98+.019*x[3])
                latency=(20+6000*x[4])*1.468*1.3/299.792458
            elif name=='SACP':
                cost,v=sacp_cost(180+420*x[0],5+15*x[1],.3+.5*x[2],.90+.08*x[3])
                latency=(5+2000*x[4])*1.468/299.792458+22/299.792458
            else:
```

```
                cost=orbital_design(T=330+60*x[0],pv_areal_mass=.8+2.2*x[1],rad_areal_mass=1+5*x[2],other_mass=2+6*
            x[3],launch=50+650*x[5],qualification=1.5+3.5*x[6],availability=.90+.09*x[7])
                latency=(5+2000*x[4])*1.468/299.792458+(500+1500*x[8])/299.792458
            rows.append([name,k,latency,cost['usd_per_EFLOP']])
    f,a=plt.subplots(figsize=(9,5),layout='constrained')
    colors=[COLORS[x] for x in ['ground','onsite','coast','SACP','LEO']]
    names=['Grid-connected','Dedicated energy','Coastal','SACP','LEO']
    for name,col in zip(names,colors):
        vals=np.array([[r[2],r[3]] for r in rows if
            r[0]==name]);a.scatter(vals[:,0],vals[:,1],s=7,alpha=.20,color=col,label=name)
    allv=np.array([[r[2],r[3]] for r in
        rows]);mask=pareto(allv[:,0],allv[:,1]);fr=allv[mask];fr=fr[np.argsort(fr[:,0])]
    a.plot(fr[:,0],fr[:,1],color='black',lw=.9,label='Frontier of sampled points only')
    a.set(xscale='log',xlabel='Client-to-compute propagation, one way (ms)',ylabel='LCOC (USD / $10^{18}$ useful
        FLOP)',ylim=(.35,2.1));grid(a);a.legend(ncol=3,fontsize=7,loc='upper right')
    save(f,'F19_exploration_cloud','F19. 2,500 exploratory configurations (500 per family, stratified Latin
        hypercube sampling, seed 2030). Densities and frequencies are neither probabilities nor observations. Local
         solar costs include assumed firm energy in their price. Telecommunications paths are geometric scenarios.
        The frontier applies to this sample, without a guarantee of global optimality or feasibility in 2030.')
    with (DATA/'F19_exploration_cloud.csv').open('w') as out:
            w=csv.writer(out);w.writerow(['architecture','sample','one_way_ms','USD_per_1e18_useful_FLOP']);w.write
            rows(rows)
    dump('F19_sampling_domains.json',dict(seed=2030,n_per_architecture=n,ground=dict(price_MWh=[40,200],infra_MUSD_
        MW=[5,15],pue=[1.10,1.30],availability=[.98,.999]),onsite=dict(price_firmed_MWh=[60,210],compute_infra=[6,1
        6]),coast=dict(price_MWh=[55,180],infra=[8,18],pue=[1.08,1.23]),SACP=dict(eb_Whkg=[180,600],wind_ms=[5,20],
        env_kgm2=[.3,.8],availability=[.90,.98],annual_opex_MUSD=2),LEO=dict(T_K=[330,390],PV_kgm2=[.8,3],rad_kgm2=
        [1,6],other_kgkW=[2,8],launch_USDkg=[50,700],qualification_MUSD=[1.5,5],availability=[.90,.99])))
if __name__ == '__main__':
    validate()
    thermal()
    deployment()
    scatter()
    print('Three benchmark figures and data written to', ROOT)
```

Principal References Cited

The following links identify the documents cited in the main text. The CSV ledgers preserve additional references and extracted values. Documentary evidence is current as of September 25, 2026.

- IEA, April 2026.
- LBNL, pages 8–10.
- IEA, February.
- IEA, July.
- IEA, Supply.
- PJM, January 2026.
- ERCOT.
- AEMO/Oxford.
- RTE, electricity review.
- RTE, data centers.
- IEA, Grids.
- IEA, equipment.
- LBNL, Queued Up 2026.
- Fingrid, August 2026.

- Svenska kraftnät, northern development.
- Statnett, Rana substation.
- Landsvirkjun, Akureyri PPA.
- Farice, AUÐUR project.
- Hydro-Québec, tariff proposal.
- Talen, SEC-filed press release.
- PJM, adequacy and large loads.
- ERCOT, application status.
- ERCOT, Batch Zero notice.
- Fervo, Cape Station.
- Fervo, 396 MW PPA.
- Google, Nevada.
- Grant PUD, Quincy grid.
- Grant PUD, 2026 IRP.
- Red Eléctrica, demand access capacity.
- MinCiencia, PDATA.
- Subtel, Humboldt Connect.
- AEMO, 2026 ISP.
- Masdar, solar and storage project.
- IMDA, Green Data Centre Roadmap.
- Guizhou government, interconnection node.
- Google, Hamina.
- Lienhard, flat-plate heat-transfer correlations.
- NASA, small-satellite thermal control.
- NASA, solar irradiance.
- U.S. Standard Atmosphere, 1976.
- USGS estimates 2025 global helium production at 190 million cubic meters.
- SpaceX, SEC filing.
- SpaceX, AI1.
- Google, Suncatcher preprint.
- Gaalema, Indyk and Staley, 2026.
- NASA, state of the art in power subsystems.
- FAA, Boca Chica.
- FAA, Kennedy study.

- NOAA and associated PNAS publication.
- World Bank, financing structures.
- World Bank, DSCR.
- Microsoft Research, Project Natick.
- NASA, TDRS systems.
- NASA lunar data.
- DSCOVR mission.
- NASA work on the lunar south pole.