

The Compute Escape

Why AI is fleeing the grid for homes, oceans and orbit, and why most escape routes fail the full-cost test

A RefleXio field guide to the infrastructure ideas that sound like science fiction, the constraints they actually solve and the costs they quietly move somewhere else.

STRATEGIC DOSSIER

Strategic Dossier No. 05

The Compute Escape

Why AI is fleeing the grid for homes, oceans and orbit, and why most escape routes fail the full-cost test

A RefleXio field guide to the infrastructure ideas that sound like science fiction, the constraints they actually solve and the costs they quietly move somewhere else.

Publication date	Late August 2026
Publisher	RefleXio Intelligence
Series	Strategic Dossiers, No. 05
Version	1.0
Prepared by	RefleXio Intelligence Editorial Desk

Independent strategic intelligence on sovereign cloud, AI compute and the European digital stack.

THE REFLEXIO THESIS. AI compute is not escaping infrastructure. It is escaping the places where infrastructure costs are easy to see. Homes, oceans and orbit can solve real constraints for narrow workloads, but none repeals the grid, maintenance, thermodynamics or finance. The winning architecture will be a routed portfolio, measured by full-system cost per accepted task, not a beauty contest between exotic locations.



Contents

Contents	2
How to read the evidence	5
Executive summary.....	6
The ten conclusions that matter.....	6
Numbers worth carrying into the debate	7
1. The crisis is real. The escape story is optional.....	8
1.1 Capital is not waiting for consensus	9
1.2 The four bottlenecks hiding inside one crisis	10
1.3 The contrarian point.....	10
2. The boundary is the business model	10
2.1 A full-system denominator	11
2.2 Metrics that belong beside PUE	12
2.3 Evidence discipline is part of the metric.....	12
3. Land is constrained. It is not obsolete.....	12
3.1 The terrestrial improvements that deserve more attention.....	13
3.2 The flexibility myth needs a contract.....	13
3.3 Land wins by becoming less monolithic.....	13
4. Residential compute or the data center behind the boiler home	14
4.1 The economic case	15
4.2 The fleet problem.....	15
4.3 Workloads that fit and workloads that do not.....	15
4.4 Residential verdict	16
5. Subsea compute or how the ocean is a heat sink, not an alibi.....	16
5.1 What the ocean genuinely improves.....	17
5.2 What the ocean makes worse	17
5.3 The environmental boundary.....	18
5.4 Commercial shape.....	18
5.5 Subsea verdict.....	18
6. Orbital computing and how solar abundance meets thermal tyranny	19
6.1 The radiator calculation	20
6.2 Launch economics.....	21
6.3 Radiation, repair and upgrade cadence.....	21
6.4 The million-satellite filing	21
6.5 Orbital verdict	22
7. The trivector showdown is the wrong contest	22

7.1 A comparison without fictional precision	23
7.2 The economic model that should replace location hype	24
7.3 The cost categories most often omitted	24
7.4 The decisive variable is workload value density	25
8. Workload routing is the architecture.....	25
8.1 Frontier training	26
8.2 Interactive inference	26
8.3 Batch and flexible inference	26
8.4 Data born in space	26
8.5 A routing policy for 2026	27
9. Finance. Exotic infrastructure is a duration trade	27
9.1 The revenue problem arrives before the engineering problem	27
9.2 Tail risk by substrate.....	28
9.3 The stranded-compute risk	28
9.4 Financing rule	28
10. Externalities do not disappear. They migrate	29
10.1 Energy and carbon	29
10.2 Water and heat	29
10.3 Material and spatial burden	29
10.4 Shared environments and governance	30
11. Security and sovereignty. And why location is just one aspect.	30
11.1 Security properties change by substrate.....	30
11.2 Isolation can reduce visibility	31
11.3 Sovereign procurement should buy switching power.....	31
12. The project map. Reality by evidence grade.	31
12.1 Selected project scorecard	32
12.2 What would move a project up one grade	33
12.3 The communication test	34
13. Europe. The advantage is at the boundary.	34
13.1 The energy contradiction	35
13.2 Heat reuse is Europe's most practical differentiator.....	35
13.3 Marine engineering is an exportable layer	36
13.4 Space should begin with space data	36
13.5 Regulation. Guardrail or ankle weight.....	36
13.6 What Europe should build by 2030	37
13.7 Europe's uncomfortable advantage	37
14. Decision-Making Framework. How to avoid buying a single location.	37

14.1 Gate one. The task	38
14.2 Gate two. The unique constraint.....	38
14.3 Gate three. The complete boundary	38
14.4 Gate four. Failure and refresh.....	38
14.5 Gate five. Evidence and externalities	38
14.6 Gate six. Portfolio value.....	39
14.7 Board checklist	39
15. Playbooks by actor.....	40
15.1 For compute operators	40
15.2 For enterprise and sovereign buyers	40
15.3 For investors and lenders	40
15.4 For policymakers	41
15.5 For communities and infrastructure hosts.....	41
16. Scenarios through 2035.....	41
16.1 Scenario A. Land adapts faster than the escape routes	41
16.2 Scenario B. Distributed edge becomes an energy appliance class	42
16.3 Scenario C. Subsea becomes coastal sovereign infrastructure.....	42
16.4 Scenario D. Orbital data processing scales, orbital cloud does not	42
16.5 Scenario E. The capital cycle disappoints before the substrates mature	42
16.6 Scenario indicators	43
17. Watch list. Evidence that can change the thesis	43
17.1 Land and grid	43
17.2 Residential.....	43
17.3 Subsea	44
17.4 Orbit.....	44
17.5 Europe.....	44
18. Conclusion. The future has an electrician.	44
19. Methodology and calculation notes.....	45
19.1 Research method	45
19.2 Electricity figures	45
19.3 Hyperscaler capital expenditure	46
19.4 Orbital radiator model	46
19.5 Solar collection illustration	46
19.6 Evidence-grade limitations.....	46
20. References	47
Disclaimer	49

READING MAP

Part	Question
I. Reality check	What is actually breaking in the terrestrial model and which metrics conceal it
II. Escape routes	What residential, subsea and orbital systems solve, and what they merely relocate
III. Portfolio logic	Which workloads belong where, under what economics and evidence threshold
IV. Europe	Where Europe has structural advantages, and where regulation becomes self-harm
V. Action	How boards, operators, investors, and policymakers can decide without buying theater

How to read the evidence

This report separates observation from promotion. A project may be physically operating while its efficiency, availability and economics remain known only through the operator. A regulatory filing is evidence of intent, not evidence of an authorized constellation. A feasibility study is useful, but it is not capacity. The distinction sounds elementary. In emerging infrastructure, it is practically insurgent.

Four grades are used throughout. Grade A is an operating fleet or public statistic supported by authoritative and sufficiently transparent evidence. Grade B is an operating asset whose performance data are still predominantly supplied by its operator or a state source. Grade C is a prototype or pilot that proves a mechanism but not commercial scale. Grade D is an announced product, filing, call for proposals or feasibility concept. The grades measure evidence maturity, not the intelligence of the engineering team or the probability that a project will eventually work.

A	Verified operating evidence. A measured fleet, audited statistic or established infrastructure class with transparent methods.
B	Operating with incomplete independent validation. Physical operation is documented, while the decisive performance claims remain operator-led.
C	Prototype evidence. The mechanism works at pilot scale. Service economics, maintainability, and scale remain open.
D	Plan or proposition. An announcement, filing, study, design target or funding call. Useful as a signal, not as installed capacity.

Executive summary

The terrestrial data center is not dying. It is becoming politically, electrically and physically visible. That visibility has created an industry of escape stories. Put servers in homes and call the electricity distributed. Put them underwater and call cooling free. Put them in orbit and call solar energy abundant. Each claim contains a serious engineering insight. Each becomes misleading when its boundary is drawn one meter too tightly.

AI infrastructure does not escape constraints. It reroutes them.

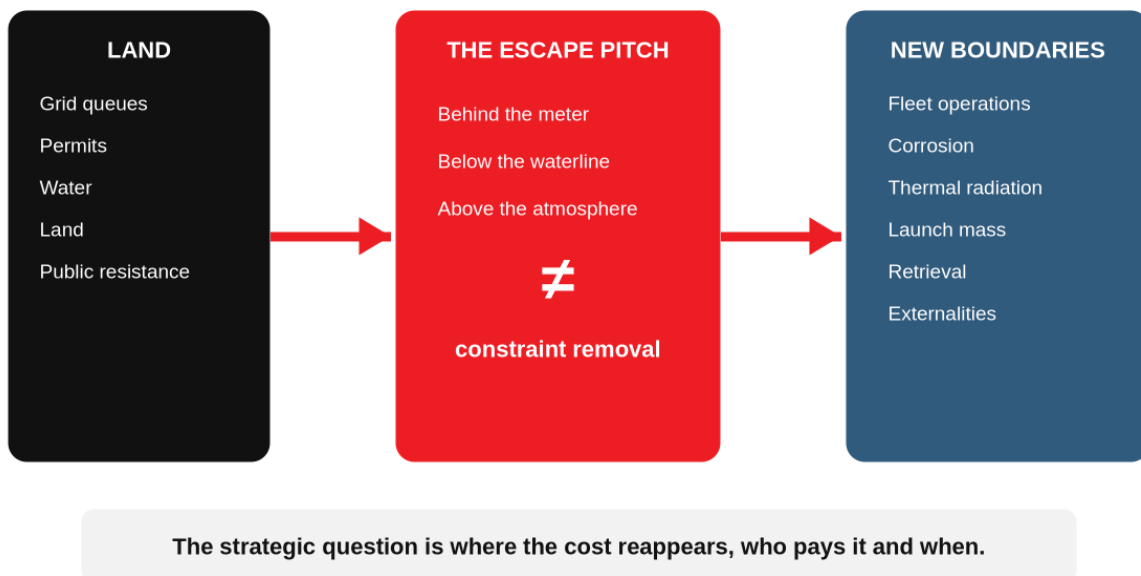


Figure 1. Constraint routing. RefleXio synthesis of the infrastructure boundary problem.

The ten conclusions that matter

1. Land remains the default substrate through 2035 because dense networking, rapid repair, financing, supply chains, and upgrade cadence still dominate most workloads. Its constraint is coordination, not technical extinction.
2. Global data center electricity use rose to about 485 TWh in 2025 and the IEA expects about 950 TWh in 2030. The question is no longer whether the load matters. The question is whether it can arrive where generation, networks and permits can arrive with it.
3. The four largest United States hyperscalers guided to roughly \$695 billion to \$720 billion of 2026 capital expenditure in aggregate. Not all of it is AI or data centers, but the figure reveals the balance-sheet scale against which exotic infrastructure must compete.
4. PUE is necessary and insufficient. It says how much facility energy accompanies IT energy. It does not price grid upgrades, launch mass, embodied equipment, ecological heat, data movement, retrieval, insurance, security or decommissioning.
5. Residential compute can turn unavoidable server heat into useful hot water and can deploy behind existing meters. It also creates a dispersed hardware fleet with household duty cycles, cyber boundaries, noise, service, insurance, and heat-demand mismatch.

6. Subsea compute has a credible physical advantage in stable cooling and reduced direct freshwater use. It pays for that advantage with corrosion control, marine permitting, cable dependence, retrieval logistics, hardware lock-in, and a thin independent ecological evidence base.
7. Orbital compute is credible first where data are already in orbit and downlink is the bottleneck. It is not credible as a general substitute for terrestrial training while launch, radiator, shielding, service, and replacement economics remain unresolved.
8. Vacuum is not a coolant. A one-megawatt payload rejecting heat at 300 K requires about 2,419 square meters of ideal one-sided radiator area at an assumed emissivity of 0.90 under a deliberately generous lower-bound model. The hardware becomes more compact only by operating hotter, which shifts the burden to components and reliability.
9. Europe is not condemned to be a customer. Its best position is in full-system orchestration, heat reuse, low-carbon siting, industrial edge, marine engineering, environmental measurement, power electronics, space robotics, and lifecycle standards. Its worst position is a permit queue decorated with sovereignty language.
10. The winning architecture is a portfolio. Route each task to the smallest substrate that satisfies latency, synchronization, security, repairability, and energy constraints. The future is less likely to be one spectacular data center than a disciplined scheduler with an electrician on call.

Numbers worth carrying into the debate

Signal	Best current figure	What it does and does not mean	Source
Global electricity	485 TWh in 2025, about 950 TWh in 2030	IEA system estimate. Includes all data center workloads, not only generative AI.	OFFICIAL IEA Energy and AI 2025: Key Questions on Energy and AI 2026 STATISTIC
United States 2030 range	521 to 843 TWh, 649 TWh reference	LBNL projection. The reference equals 11.8% of United States electricity.	OFFICIAL LBNL US data center energy usage report, Jun 2026 STATISTIC
Europe	87 TWh in 2024, 134 TWh in 2030	ENTSO-E estimate. More than 10,500 facilities above 50 kW are counted.	OFFICIAL ENTSO-E, Data centers and electricity systems, May 2026 STATISTIC
Hyperscaler capex	\$695B to \$720B guided for 2026	Alphabet, Amazon, Meta, and Microsoft aggregate. Includes infrastructure beyond AI.	COMPANY Alphabet, Amazon, Meta and Microsoft 2026 guidance; RefleXio aggregation FILING
Subsea project scale	Shanghai Lingang design at 24 MW	Operating project reported by official sources. Efficiency claims remain operator or state led.	OPERATOR OR STATE LED Lingang Special Area Administration, Aug 2025; Xinhua, Mar 2026
Orbital filing	Up to 1,000,000 satellites	SpaceX application accepted for filing by the FCC. It is not an authorization or deployed system.	COMPANY FCC acceptance for filing, Feb 2026 (SpaceX requested figure) CLAIM

Signal	Best current figure	What it does and does not mean	Source
Radiator lower bound	2,419 m ² per MW at 300 K, emissivity 0.90	RefleXio calculation. Excludes solar, Earth infrared, structure, redundancy, and pumping.	REFLEXIO JUDGMENT RefleXio Stefan-Boltzmann calculation, methodology 19.4
European AI build	Up to 7 gigafactories, more than €30B intended	European Commission call. It is procurement and financing intent, not connected capacity.	OFFICIAL STATISTIC European Commission, AI gigafactories call, Jul 2026

Table 2. Numbers worth carrying into the debate.

Source note. IEA, 2025, 2026; Lawrence Berkeley National Laboratory, 2026; ENTSO-E, 2026; Alphabet, Amazon, Meta, and Microsoft investor disclosures, 2026; FCC, 2026; RefleXio calculation.

1. The crisis is real. The escape story is optional.

The infrastructure problem has crossed from industry planning into macroeconomic relevance. The International Energy Agency estimates that data centers used about 485 TWh in 2025, up 17% in one year, and projects roughly 950 TWh by 2030. AI-focused centers grew even faster, and their consumption is expected to triple between 2025 and 2030. A rack built around the most advanced accelerators in 2027 could draw peak power comparable to about 65 households. Rack density increased roughly elevenfold between 2020 and 2025 and may increase another fourfold by 2027 (International Energy Agency [IEA], 2026).

Those numbers justify urgency. They do not justify indiscriminate panic. Data center demand is concentrated, schedulable to varying degrees and often associated with companies able to finance generation, transmission and storage. The same IEA analysis notes that data centers can host 20 to 25 GW of batteries by 2030 and that price effects depend heavily on whether load and investment are coordinated. The infrastructure crisis is therefore partly a coordination crisis. Electrons exist. The difficult work is connecting generation, wires, land, equipment, contracts, and political permission on the same calendar.

The United States illustrates the range of uncertainty. Lawrence Berkeley National Laboratory projects 649 TWh of data center electricity use in 2030 in its reference case, with a range of 521 to 843 TWh. The reference equals 11.8% of national electricity. Five years earlier, such a share would have sounded like a rounding error in a bad science-fiction script. It now sits in an official model with sensitivity bands (Lawrence Berkeley National Laboratory [LBNL], 2026).

Europe starts from a smaller load and a more fragmented energy map. ENTSO-E estimates that European data centers consumed 87 TWh in 2024 and could reach 134 TWh in 2030. Its 2035 range reaches 199 to 254 TWh. It also counts more than 10,500 facilities above 50 kW and estimates approximately 12.7 GW of installed IT power across Europe. The headline is growth. The deeper finding is spatial friction. New connections can take from several years to more than a decade, and flexibility that looks generous in a model is heavily filtered by contractual, operational, and technical constraints in practice (European Network of Transmission System Operators for Electricity [ENTSO-E], 2026).

This is the context in which escape architectures appear rational. A distributed home node avoids a campus permit. A subsea module avoids cooling towers and expensive coastal land. An orbital platform points at continuous solar generation and bypasses terrestrial connection queues. The pitch is not

foolish. The mistake is to compare the removed constraint with nothing. Every substrate has a balance sheet, a failure mode and a neighbor, even if the neighbor is a fish or a satellite.

Data center electricity moves toward a one-petawatt-hour year

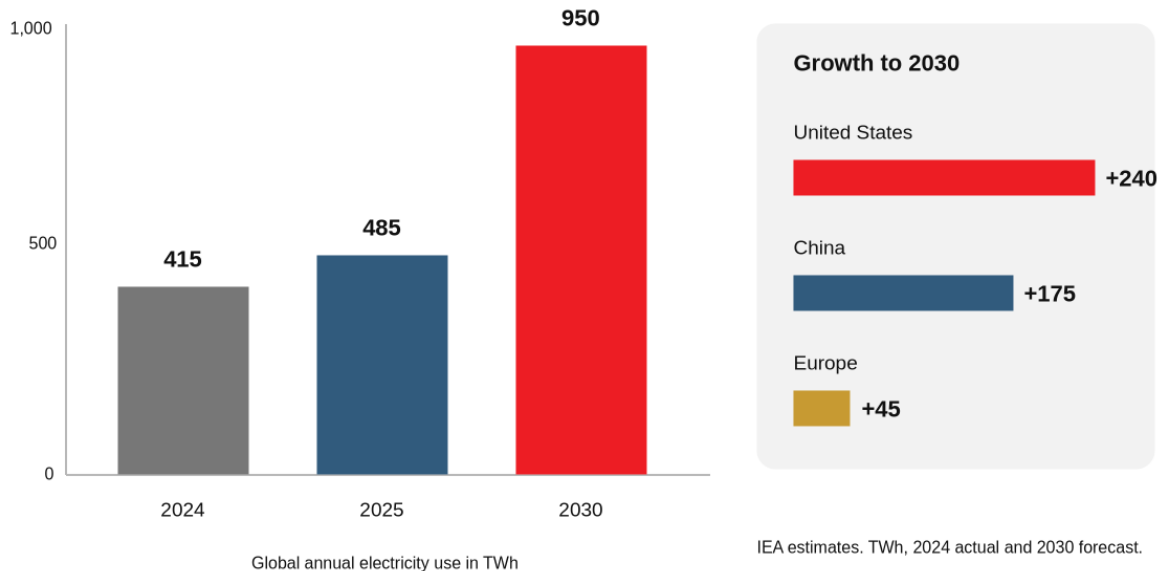


Figure 2. Global data center electricity and regional growth. IEA 2025 and 2026 estimates.

1.1 Capital is not waiting for consensus

The infrastructure race is already capitalized at a scale that changes the competitive test. Alphabet guided to \$175 billion to \$185 billion of 2026 capital expenditure. Amazon guided to approximately \$200 billion. Meta guided to \$130 billion to \$145 billion. Microsoft indicated roughly \$190 billion for calendar 2026. The combined range is approximately \$695 billion to \$720 billion (Alphabet, 2026; Amazon.com, 2026; Meta Platforms, 2026; Microsoft, 2026).

The aggregation must be handled carefully. These budgets include facilities, servers, networks, chips, robotics, logistics, and other infrastructure. They are not a clean measure of AI data centers. They are still strategically decisive. A new substrate does not compete against an abstract server room. It competes against incumbents that can buy generation, negotiate transmission, design custom silicon, prepay suppliers, absorb utilization troughs, and refinance failures. An elegant module with a beautiful thermal diagram is not yet an infrastructure business.

This imbalance explains why many exotic projects begin as complements. Residential nodes seek capacity that can be installed faster. Subsea modules target coastal or sovereign niches. Orbital nodes preprocess data that would otherwise consume scarce downlink. The modest versions can be coherent. The grand versions often require the new substrate to beat land on every variable at once. Engineering rarely offers that kind of kindness.

1.2 The four bottlenecks hiding inside one crisis

Bottleneck	Visible symptom	What actually relieves it	What merely moves it
Power	Connection queue, curtailment risk, high local prices	New generation, wires, storage, flexible operations, credible contracts	Moving behind a meter without adding system capacity
Cooling	Water limits, fan and chiller energy, hot racks	Liquid cooling, heat reuse, better siting, higher-temperature components	Rejecting heat into another environment without measuring the plume
Land and permission	Permit delay, community resistance, grid congestion	Better planning, benefit sharing, industrial siting, modular approval	Hiding the asset offshore or inside homes while retaining public dependencies
Compute economics	Low utilization, rapid obsolescence, poor margins	Workload matching, software efficiency, procurement discipline	Assuming cheap energy cures idle or stranded hardware

Table 3. The four bottlenecks hiding inside one crisis.

1.3 The contrarian point

REFLEXIO VIEW. The most important scarce resource is not electricity by itself. It is coordinated, financeable and socially permitted electricity delivered on the same schedule as compute hardware. Exotic substrates are valuable when they reduce that coordination cost for a specific workload. They become theater when they count only the constraint they remove.

2. The boundary is the business model

Power usage effectiveness, or PUE, is the ratio of total data center energy to IT equipment energy. It remains a useful facility metric, and the industry is right to use it. The Green Grid created it to reveal overhead. ISO/IEC 30134-2:2026 codifies it within a suite of resource-efficiency metrics and explicitly notes that a holistic approach requires the suite, not one ratio in isolation (International Organization for Standardization, 2026; The Green Grid, n.d.).

The trouble begins when PUE is promoted from facility ratio to planetary verdict. Google reports a 2025 fleet PUE of 1.09, an impressive operating result (Google, 2026a). Yet even a hypothetical PUE of 1.05 says nothing about whether a new transmission line was socially useful, whether a gas turbine was built for a low-utilization campus, whether water was consumed in a stressed basin, whether heat was dumped into a marine habitat, whether hardware had to be launched, or whether the answer crossed a network three times before reaching the user.

The correct comparison is not server watt against server watt. It is full-system cost per accepted compute task under a defined service level. Accepted matters because a fast wrong answer can be cheap only in the accounting department. Service level matters because delay-tolerant batch inference,

frontier training and safety-critical control are different products. Full system matters because infrastructure has an inconvenient habit of sending invoices after the pitch deck has closed.

PUE can improve while the full system gets worse

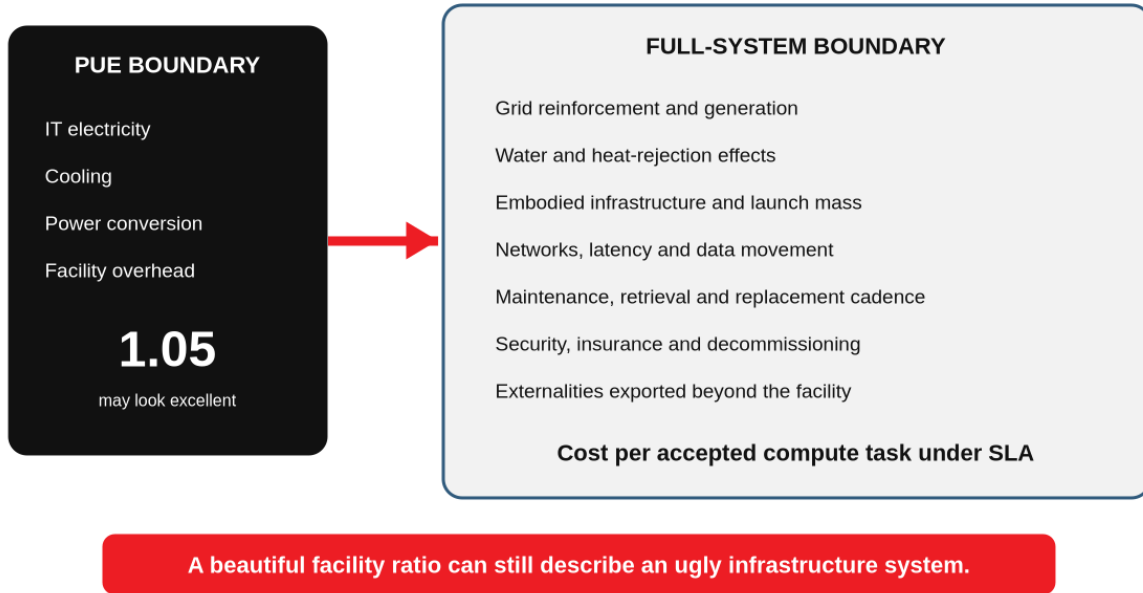


Figure 3. The full-system boundary. RefleXio framework extending beyond facility PUE.

2.1 A full-system denominator

RefleXio proposes a decision denominator rather than a new universal standard. The unit is the expected full-system cost of an accepted compute task delivered under its latency, availability, security, and sovereignty requirements. It includes facility energy, external power infrastructure, water, heat rejection, embodied equipment, network movement, maintenance, replacement, insurance, security, decommissioning, and monetized externalities where credible.

The denominator can be expressed conceptually as follows. Full-system task cost equals annualized infrastructure and energy cost, plus expected failure and externality cost, divided by accepted tasks delivered within the required service level. The equation is deliberately boring. Boring is useful. It prevents a project from claiming victory by exporting its hardest variable outside the diagram.

This framework also explains why no substrate wins universally. A terrestrial campus may have higher local water use but lower network, repair and replacement cost. A home node may create valuable heat but lose utilization when the household does not need it. A subsea module may achieve stable thermal behavior but require expensive retrieval for one failed power interface. An orbital node may save downlink bandwidth by preprocessing satellite imagery while failing completely for a conversational workload whose users and data are on Earth.

2.2 Metrics that belong beside PUE

Metric	Decision question	Why it changes the answer
Accepted-task energy	How many joules produce a usable result under SLA	Captures software, retries, routing, and model quality.
Grid additionality	What generation, storage and wires exist because of the load	Separates procurement claims from physical capacity.
Useful heat ratio	What share of waste heat displaces another energy service	Credits real heat reuse only when timing and temperature match demand.
Service intervention rate	How often hardware needs physical action	Makes distributed, subsea and orbital maintenance comparable.
Upgrade half-life	How long the installed generation remains economically useful	Prices hardware lock-in and accelerator obsolescence.
Network work	How much data moves, how far and with what latency	Prevents location from disappearing from compute economics.
Boundary externality	Which costs are exported to communities, oceans or orbit	Reveals who pays outside the facility invoice.

Table 4. Metrics that belong beside PUE.

2.3 Evidence discipline is part of the metric

A full-system model can still become decorative if every unknown receives an optimistic value. This report therefore distinguishes measured values, operator claims, design targets, and RefleXio calculations. Microsoft measured Project Natick against a land replica. The Shanghai subsea project publishes design and operating claims through official channels. SpaceX filed for a satellite system. NVIDIA announced orbital modules. These are different evidence events and must not share a tense.

The rule is simple. Operating is not the same as commercial. Launched is not the same as utilized. Authorized is not the same as filed. A design power is not an average load. A target PUE is not a measured annual PUE. If a report cannot preserve those distinctions, its precision is cosmetic.

3. Land is constrained. It is not obsolete

Land keeps several unfair advantages. Engineers can reach it. Fiber can be repaired without a ship or launch window. Failed accelerators can be replaced on the cadence of the semiconductor market. Dense clusters can synchronize through high-bandwidth fabrics. Insurers and lenders understand the asset. Suppliers know how to deliver it. Users are usually on the same planet, which remains a surprisingly effective latency optimization.

These advantages are why terrestrial capacity keeps attracting extraordinary capital despite grid queues. The dominant response is not abandonment. It is vertical integration. Operators secure energy

contracts, build substations, procure transformers, design chips, deploy batteries, and consider on-site generation. The IEA estimates that data centers could operate 20 to 25 GW of battery storage by 2030. It also estimates that United States on-site gas generation for data centers could reach 15 to 27 GW by 2030, with 30% to 70% overbuild relative to data center demand if the facilities are to maintain reliability during outages (IEA, 2026).

That last figure exposes a hard truth. The escape from the grid can reproduce the grid badly. A self-supplied campus still needs redundancy, fuel logistics, emissions controls, maintenance, and often a grid connection. Behind-the-fence generation can accelerate deployment, but it can also create underused parallel infrastructure whose costs are socialized through fuel markets or local air quality. The architecture should be judged by system value, not by whether the utility meter is visible.

3.1 The terrestrial improvements that deserve more attention

1. Energy-aware scheduling that moves delay-tolerant work across hours or regions without violating data and latency requirements.
2. High-temperature liquid cooling that increases heat quality and makes district-heating reuse less absurdly seasonal.
3. Modular substations, standardized permits and transmission planning that align connection dates with equipment delivery.
4. Hardware and model efficiency that reduce joules per accepted task, including quantization, sparsity, caching, routing, and smaller specialized models.
5. Flexible battery and backup assets that provide grid services instead of sitting idle as expensive reassurance.
6. Industrial and brownfield siting where power, heat customers, transport, and community benefit can be planned together.

3.2 The flexibility myth needs a contract

Data centers are often described as flexible loads. Some are. Training runs can sometimes pause. Batch inference can move. Batteries can respond quickly. Yet interactive services, synchronization-heavy jobs, safety workloads, and customer contracts reduce the technically available flexibility. ENTSO-E warns that the theoretical potential is heavily filtered in practice. It also notes a different system risk. Very large facilities can disconnect hundreds of megawatts almost instantaneously during a disturbance, turning a load problem into a balancing problem (ENTSO-E, 2026).

Flexibility should therefore be procured as a measured product. Contracts should specify power, duration, response time, recovery behavior, workload class, and compensation. An operator that advertises flexibility but cannot expose an auditable dispatch window is offering a mood, not a grid service.

3.3 Land wins by becoming less monolithic

The future terrestrial system is likely to be layered. Frontier training remains in dense campuses. Interactive inference moves closer to users. Stable batch jobs move toward low-cost, low-carbon power. Industrial workloads remain near data and process equipment. Heat-reuse nodes sit near reliable heat demand. The relevant unit is no longer the data center. It is the routed workload portfolio.

This shift matters because it makes exotic substrates less revolutionary and more useful. A home node, a subsea module and an orbital processor become specialized edges within a larger terrestrial control

plane. Their value rises when the task is narrow. Their credibility falls when they claim to replace the control plane itself.

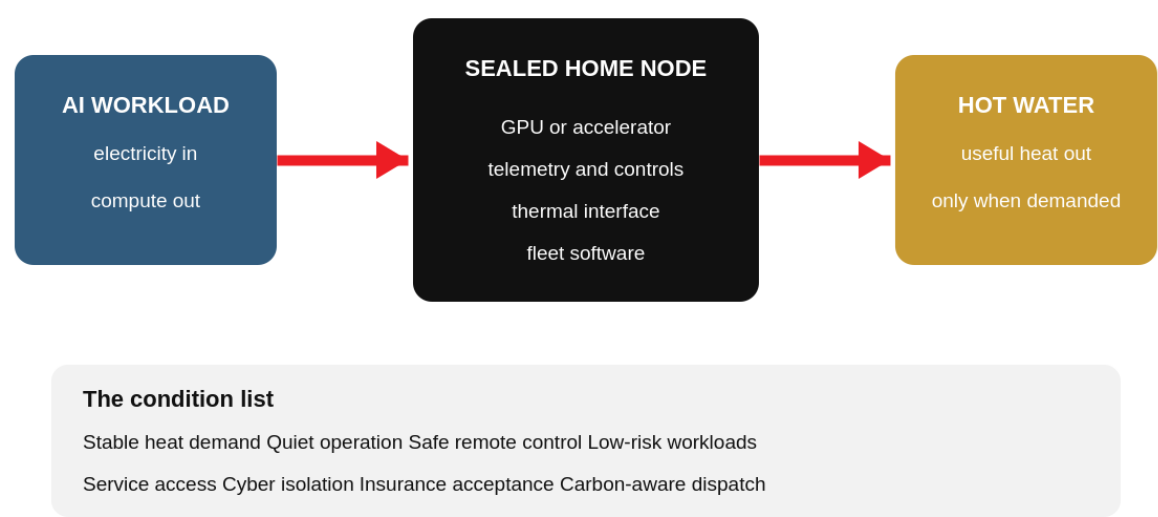
4. Residential compute or the data center behind the boiler home

Residential compute is the least cinematic escape route and may be the most immediately practical. A server consumes electricity and releases nearly all of it as heat. If that heat can displace electric resistance or gas water heating, the same input watt produces computation and a useful thermal service. The concept does not create free energy. It creates a second customer for an unavoidable output.

British Gas and heata announced a ten-unit, three-month trial in 2025. The system attaches a compact compute unit to a household hot-water cylinder and can provide up to 4 kWh of heat per day. The partners estimated annual savings of about £340 for households using electric water heating and £120 for gas-heated households. These are trial and company estimates, not a fleet-wide audited outcome (Centrica, 2025).

SPAN announced XFRA in April 2026 as a distributed data center architecture using residential and small commercial nodes with an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU. Initial deployment was announced for later in 2026, with a pipeline intended to scale toward gigawatts in 2027. No operating fleet or audited utilization was disclosed with the announcement (SPAN, 2026). The correct grade is D. Ambition is visible. Operations are not yet available for inspection.

Residential compute monetizes the same watt twice, when conditions coope



A warm shower is useful. Ten thousand unserviceable GPU boilers are a fleet-management seminar.

Figure 4. Residential heat reuse. The value proposition and the operating conditions that govern it.

4.1 The economic case

The strongest residential case has four features. The workload is stable and not extremely latency sensitive. The home has predictable hot-water demand. Electricity is low carbon or dispatched when marginal emissions are low. The node can be serviced without turning households into reluctant data center technicians.

Under those conditions, heat reuse changes the effective energy cost. A conventional data center pays for electricity and then for heat rejection. A residential node can sell compute while displacing some separate heating expenditure. The credit should equal the avoided energy service, adjusted for conversion efficiency, timing, and the carbon intensity of the displaced fuel. Crediting every thermal kilowatt-hour at retail electricity price would be enthusiastic accounting if the alternative was an efficient heat pump or low-cost district heat.

The behind-the-meter advantage is similarly conditional. A household connection may have spare capacity at some hours, but a fleet of dense nodes can still affect local transformers and feeders. Distribution grids were not designed around neighborhoods behaving like synchronized accelerator clusters. The meter can be hidden behind a kitchen wall. The electron still keeps accounts.

4.2 The fleet problem

Variable	Central facility	Residential fleet	Decision consequence
Physical access	Controlled and staffed	Thousands of private locations	Service labor and consent become core economics.
Thermal demand	Heat is rejected continuously	Heat value varies by household and season	Utilization and useful heat can diverge.
Cyber boundary	Purpose-built network and controls	Shared building and consumer environment	Isolation and remote attestation are mandatory.
Hardware life	Planned refresh and parts inventory	Dispersed replacement and return logistics	Obsolescence becomes a reverse-logistics problem.
Acoustics and safety	Industrial specification	Domestic expectations	Noise, fire, water, and insurance standards tighten.
Grid behavior	Known point load	Aggregated distribution load	Fleet dispatch must respect local network constraints.

Table 5. The fleet problem.

4.3 Workloads that fit and workloads that do not

Residential nodes favor inference, rendering, compilation, transcoding, and other divisible jobs that tolerate intermittent capacity. They are poor homes for tightly synchronized frontier training, highly sensitive data without strong isolation, or services requiring four-nines availability from each device. The fleet can provide reliability through redundancy, but redundancy is not free. It increases idle capacity, network traffic, and operational complexity.

The most credible near-term model is not a random marketplace of household GPUs. It is a managed appliance with a narrow hardware bill, remote attestation, secure boot, automated workload isolation, thermal integration, explicit service rights, and an aggregator that can curtail by distribution zone. This looks less like a crypto-mining revival and more like an energy appliance company that happens to sell computation. That is a compliment. Infrastructure improves when it stops dressing like a token.

4.4 Residential verdict

REFLEXIO VERDICT. Residential compute is a credible niche for stable, low-risk workloads near useful heat demand. Its moat is not cheap GPUs. It is fleet operations across energy, appliances, cybersecurity, insurance, and service. The category becomes attractive when the heat credit is measured honestly and the node can disappear from the resident's life.

5. Subsea compute or how the ocean is a heat sink, not an alibi

Subsea compute begins with a strong physical proposition. Seawater offers stable temperatures and enormous heat capacity. A sealed module can reduce humidity, dust, oxygen, and human interference. Coastal deployment can place compute near population centers without consuming prime land. Direct freshwater use for cooling can fall sharply. These are not fantasies. Microsoft proved several of them with Project Natick.

The second Natick phase placed a 240 kW module with 864 servers and 27.6 petabytes of storage off the Orkney Islands. It operated underwater for roughly two years. Microsoft reported a server failure rate one eighth that of a land-based control group and attributed the result to a stable nitrogen atmosphere and the absence of human disturbance. The project was designed for deployments lasting up to five years. It was a successful research prototype, not a commercial fleet (Microsoft, 2020, n.d.).

China moved the concept toward larger operating projects. Official Lingang sources describe a 24 MW Shanghai project with investment of CNY 1.6 billion, a 2.3 MW first phase, designed PUE below 1.15 and more than 90% green electricity. Chinese government sources reported completion in October 2025 and state media reported operation beginning in February 2026. The same sources claim 22.8% lower power use, elimination of cooling-water consumption and more than 90% lower land use. Those figures are useful disclosures. They are not equivalent to an independent annual performance audit (Lingang Special Area Administration, 2025; State Council of the People's Republic of China, 2025; Xinhua, 2026).

Hainan's underwater center has reportedly operated since 2022. Xinhua reported 675 PFLOPS of computing power and zero server failures in the deployed modules as of mid-2025. Again, the physical operation is meaningful and the performance is an operator or state-reported claim. A credible analyst can be impressed and still ask for the maintenance log (Xinhua, 2025).

Subsea compute exchanges water and land scarcity for retrieval and ecology

The ocean is an excellent heat sink. It is also an operating environment, not a free utility.

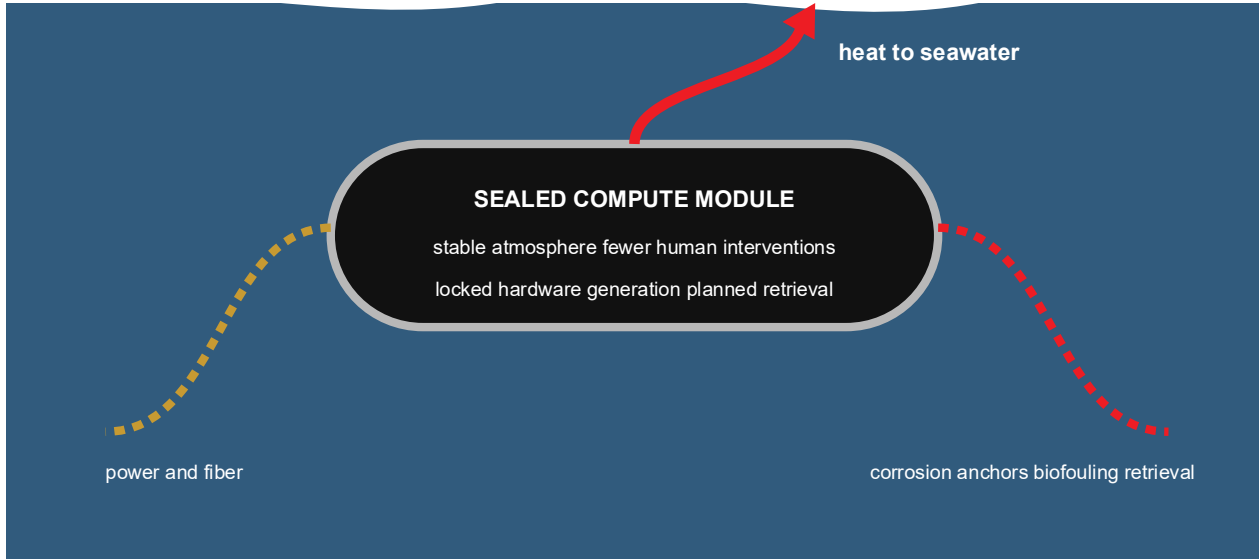


Figure 5. Subsea trade-off. Cooling and land advantages sit beside retrieval, corrosion and ecological duties.

5.1 What the ocean genuinely improves

1. Stable ambient temperatures can simplify heat rejection and reduce compressor or fan energy.
2. Sealed modules can reduce particulate contamination, humidity variation and accidental human interference.
3. Coastal proximity can reduce user latency relative to remote inland campuses, if fiber routes are well designed.
4. Direct use of freshwater for cooling can approach zero at the module, though the full electricity supply still has water and environmental implications.
5. Factory-built modules can support repeatable commissioning and rapid placement once marine permits, cables and landing infrastructure exist.
6. Physical isolation can strengthen some security properties while making recovery and inspection more difficult.

5.2 What the ocean makes worse

The basic maintenance bargain is brutal. Land systems are designed around component replacement. A subsea system is designed around failure containment until scheduled retrieval. Better server reliability helps, but the decisive risks may sit outside the server. Power connectors, fiber interfaces, pressure boundaries, pumps, moorings, and corrosion protection can determine the fate of the whole module. A cheap failed component can trigger an expensive ship operation.

Hardware obsolescence is another hidden tax. Accelerator generations move faster than subsea deployment and retrieval cycles. A module designed for five years may remain technically alive while its computation becomes economically old. The right workload must therefore value stability over

frontier performance. Storage, sovereign batch processing, deterministic inference, and long-lived services fit better than a race for each new accelerator generation.

Marine heat is the least mature part of the public evidence. A 2023 meta-analysis of thermal pollution from coastal nuclear power plants found an average local seawater temperature increase of 4.38 degrees Celsius near discharge points. The study was not about underwater data centers and cannot be applied directly to a sealed diffuse module. It does establish that thermal plumes are ecological variables, not rhetorical details (Guimarães et al., 2023). Microsoft reported rapid mixing around its small prototype. Multi-megawatt commercial arrays need their own long-duration, site-specific monitoring.

5.3 The environmental boundary

Claim	Narrow truth	Full-system question
No cooling water	The module may consume no freshwater for cooling	What water and ecological effects sit in generation, manufacturing and marine heat rejection
No land	The pressure vessel occupies seabed rather than urban land	What cable landings, ports, substations, and protected zones are required
Free cooling	Cold seawater lowers heat-rejection work	What pumps, exchangers, coatings, and ecological controls are needed
More reliable	Sealed environments can reduce component failures	Can connector, cable and vessel reliability sustain the same claim at fleet scale
More secure	Physical access is difficult	Does difficult access improve security more than it worsens recovery and oversight

Table 6. The environmental boundary.

5.4 Commercial shape

A realistic subsea industry would resemble marine energy more than cloud real estate. It needs standardized pressure vessels, qualified connectors, survey vessels, cable installation, corrosion, and biofouling expertise, environmental baselines, retrieval contracts, port logistics, and decommissioning bonds. The moat is the service ecosystem, not merely the capsule.

This profile favors state-backed or infrastructure-backed projects in locations where coastal land, freshwater, security or permitting has exceptional value. It does not yet support a universal claim that underwater capacity is cheaper. Publicly comparable data on capital cost, ship interventions, insurance, outage duration, module yield, replacement cadence, and ecological impact remain too thin.

5.5 Subsea verdict

REFLEXIO VERDICT. Subsea compute has crossed from laboratory curiosity into operating infrastructure, especially in China. It remains a specialized solution whose thermal advantage is real and whose commercial evidence is incomplete. Treat every performance number with its evidence grade attached. The ocean does not sign service-level agreements.

6. Orbital computing and how solar abundance meets thermal tyranny

Orbital computing became materially less hypothetical in 2025 and 2026. Starcloud says Starcloud-1 launched in November 2025 with an NVIDIA H100 and ran model inference and a small training experiment in orbit. That establishes that a commercial accelerator can be powered, cooled, and operated on orbit. It does not establish sustained throughput, radiation-tolerant duty cycles, thermal performance over a full year, or serviceable economics, and the company has not published telemetry that would allow those to be assessed independently. Axiom Space placed a prototype data-center unit on the International Space Station in late 2025 and launched two free-flying nodes in January 2026. NVIDIA announced the Space-1 Vera Rubin Module and related space computing products in March 2026. SpaceX filed an application for a system of up to one million satellites supporting orbital data centers. The FCC accepted the application for filing in February 2026 (Axiom Space, 2026; Federal Communications Commission [FCC], 2026; NVIDIA, 2026; Starcloud, 2026a).

The sequence matters. An H100 can operate in orbit. Compute nodes can exchange data through optical links. Product vendors see a market. A major launch provider has placed a breathtaking architecture into the regulatory process. None of these facts proves that general cloud compute is economical in orbit. They prove that the engineering frontier is moving and that the market deserves adult analysis rather than laughter or worship.

The orbital pitch combines three advantages. Solar irradiance above the atmosphere is about 1,361.6 W per square meter. Space platforms can be co-located with remote-sensing and communications data. Launch economics may improve through reusable vehicles. Each advantage is real. The phrase free cooling is not. In vacuum, heat leaves mainly through radiation. Vacuum is an excellent insulator, which is awkward when the product is heat (National Aeronautics and Space Administration [NASA], 2023, 2026a).