Data Centers in Space

Alexander Gerasimov, N4A Analytics

Sergey Maltsev, independent expert in space technologies

“Life cannot be simple. Life must be lived with passion!” — S. P. Korolev

In childhood, I read, breathlessly, a science-fiction book I had stumbled upon in our home library: Andre (Andrew) Norton’s Sargasso of Space. It tells of a distant future in which humans, having reached the stars, begin to discover worlds burned to ashes in an unimaginable war that had taken place millions of years before humanity entered space. Norton’s message is clear: the more advanced our technologies become, the greater our responsibility to use them for the good of civilization rather than to its detriment. Technological progress must go hand in hand with better, more humane social relations; otherwise, it leads to disaster.

This brings to mind another quotation from Sergey Korolev, about the harmony of nature and technology: “...We must understand and study this vast, constantly changing mechanism and, without breaking it, somehow connect the machine of our civilization to it...”

Why Put Data Centers in Space?

But let us turn to technology. Data centers in space. It sounds like an April Fool’s joke and, at the same time, like another wild fantasy from trillion-dollar companies trying to push their valuations even higher. As if they had anywhere higher to go.

The public case made by the tech giants is not very convincing. Space-based data centers as a solution to a temporary energy shortage in some U.S. regions — a shortage that suddenly appeared because of the rapid construction of giant data centers for training LLMs? This recalls a line from one of Mikhail Zhvanetsky’s sketches: “It seems the collective farm is not real... Those who have never been to a collective farm do not quite believe it, while those who live there curse Hollywood with all their might.” Anyone with even a modest understanding of data-center economics is just as puzzled: how could anyone possibly sell this to the public?

If the real goals of space-based data-center constellations truly match the stated ones, the only sound strategy for outsiders is to watch from a safe distance and wait for the next technology bubble to burst. Especially since SpaceX, while preparing for its IPO, is candidly warning potential investors that the idea of mega data centers in space carries a high risk of never materializing because of technological novelty and unclear economics.

But what if that is not the case? What if data centers in space are, after all, a real need of the economy of the future — and, in a certain technical configuration, can be built for reasonable money in the not-so-distant future?

Let us begin with the economic need: the real one, not the one invented for loud media headlines or to inspire investors to overlook fantastical revenue and profit multiples. And let us look at this need together

with the technical aspects — not the space-related ones for now, but the purely IT-related ones. In other words, let us temporarily set aside the word “space” and focus on the acronym “data center”.

The economy of the future is a digital economy, technologically realized through “AI entering the physical world”, the real sector of the economy, as a means of automating production and business processes to the point of full autonomy from direct human involvement — not only in the execution of these processes, but also in their modification. This, not writing term papers and essays for students and schoolchildren, may be the core economic meaning capable of making large-scale AI projects pay off.

Five years ago, at N4A Analytics, we tried to calculate the relationship between the costs and possible economic impact of full-scale digitalization of the real sector. We did this not for an abstract “global economy” made up of very different national economies, but using two specific economies as examples: Russia and the United States. Russia served as an example of an economy with low GDP per capita — similar per-capita indicators are typical for most BRICS+ countries, including China — while the United States served as an example of a “wealthy” economy with high GDP per capita.

The results were similar. Two common aspects stood out.

Aspect No. 1: a structural imbalance. The gross economic effect of digitalization is distributed extremely unevenly across industries, and its distribution does not correspond at all to the sectoral distribution of GDP. The lion’s share of the economic effect comes from logistics and transport, where the central task of digitalization is dynamic optimization-based management of vehicles — that is, of moving objects. This applies at every level: from direct vehicle control to fleet management and even to production, trade and logistics cooperation chains. It is also worth noting that the lower the GDP per capita, the higher the share of logistics and transport in the total gross effect of digitalization.

Aspect No. 2: the cost of digitalization. Our calculations show that digitalization is an extremely expensive project. It is no coincidence that the first “run at digitalization” in the USSR in the 1960s, at the peak of the scientific and technological revolution in the defense industry, was estimated to be more costly than the Soviet rocket-and-space, nuclear, and air-defense/missile-defense programs combined. Nothing has fundamentally changed since then: the costs are indeed space-and-nuclear in scale.

Of course, whether costs are “high” or “low” is a relative matter. Here they must be compared with the economic benefit. And the problem lies precisely in this ratio: in most sectors of the real economy, the economic benefit of digitalization turns out to be either very close to the cost or even lower. There is one caveat: that is true if unit costs are taken “as they are”. Hence a simple conclusion: unit costs of digitalization must be radically reduced.

Digitalization costs have two components: ICT infrastructure and labor. Let us begin with labor, but only briefly, since this article is about ICT infrastructure.

Digitalization is extremely labor-intensive. This remains true even in the scenario we modeled, where all applications needed for corporate digitalization are provided to corporate users as ready-made aaS products, with the required SLA. The main labor intensity lies in the need for deep restructuring, or reengineering, of production and business processes, and in creating another layer of processes above them — coordination processes that cover the entire cooperation chain end to end. Most of these costs fall on the businesses being digitalized themselves; they are internal costs. Examples are easy to find: marketplaces, which are in effect technology companies. And the lower the GDP per capita, the stronger the tilt toward in-house costs in the overall cost structure of digitalization.

How can these internal unit costs be reduced? A possible answer is beginning to emerge: AI agents. In theory, using them reduces or even eliminates the need for a long and extremely labor-intensive process-reengineering effort. With AI agents, what is needed is only a starting point suitable for launch — a minimally workable process configuration — and a properly designed process for the agents’ self-learning. The same approach can be applied to reducing the labor intensity of developing packaged software for process automation at any level, from industrial control systems to ERP, BI and BPM.

Now let us turn to ICT infrastructure. It has two components of its own: computing and telecommunications.

We will start with computing. Yes, the cost of developing it is already enormous. In the United States alone, the total volume of planned construction of new data centers for AI workloads had exceeded $800 billion by the end of 2025 — and these are private investments only. In other words, they are expected to pay off. It is assumed that further growth in data-center size will significantly reduce the unit cost of computing. We are talking not merely about hyperscale facilities, but about data centers so large that it is hard to find the right word for them: entire cities.

Yet there is another critically important factor determining the unit cost of computing: the average utilization of computing capacity. This can vary widely, from single-digit CPU utilization to above 60%. Clearly, utilization of only a few percent completely destroys the economics of any data center, wherever it is located and whatever its size.

The problem of increasing average utilization of computing resources for AI workloads is now extremely acute and still very far from being solved. SpaceX, for example, is reportedly ready to pay $60 billion for the small startup Cursor so that it can help raise the utilization of SpaceX's Colossus AI supercluster — the world's largest supercomputer for AI workloads — from a disastrous 11% to at least 30-35% when training LLMs. Utilization levels of 60% CPU load and above, familiar from other types of hyperscale workloads, are not yet even part of the discussion.

The average utilization of computing capacity is mainly shaped by the quality of the data-center, or domain, management system and by the cross-domain orchestration system. For now, however, the problem is being tackled not at the level of management systems, but mostly through infrastructure — by making data centers even larger. That makes it easier to multiplex the workloads of different tenants, combining large and small useful loads. But as the SpaceX example shows, this is far from a universal recipe; it may even be a dead end. The other extreme is optimizing LLM training processes, which Cursor is doing, but without intervening at the level of distributed network-computing infrastructure management. Clearly, optimization is needed at all three levels: from AI models to infrastructure, through the intermediate layer of infrastructure management.

Let us return to computing infrastructure. The other side of the centralized hyperscale model, when applied to AI-based control of physical objects, is intensive traffic between the controlled object and the hyperscale data center, with strict requirements for channel availability and signal latency. But network resources are limited, especially when radio spectrum is involved — and we are primarily talking about controlling moving objects, because this is where the main economic effect of digitalization is formed.

The ideal ICT-infrastructure configuration is a flexible system of root and edge computing capacities, with managed network connectivity between them, able to adapt dynamically and in real time to the current load of different applications, including AI. In theory, such a system can minimize the load on the network component and reduce end-to-end latency requirements by moving the most latency-sensitive computing load closer to the controlled object. If latency of 10 ms or less is needed between the object and the computing center, this is difficult to achieve in practice without edge computing. And even then, part of the computing must be performed directly on board the controlled object.

For this configuration to be economically viable, high average utilization must be ensured not only for root but also for edge computing capacities, as well as for the communication networks between them, the controlled object, and the computing resources on board the object. This is an extremely complex task and, to put it mildly, is still not a focus of attention for the tech giants. Trying, within our modest means, to fill this gap, we have devoted many publications to the topic, the most recent of them in the March-April 2026 issue of Connect magazine. The most problematic aspect is balancing computing and network loads between domains managed by different economic actors.

Let us come back to edge computing and ask: where, exactly, is the "edge"? Where should this edge be physically located? In general, edge computing is discussed mainly in the context of 5G networks under the term MEC. Physically, this means server cabinets at 5G base-station sites.

Why only 5G? Why not LTE? Because the separation of control and user-data planes in 5G — that is, separate transmission of signaling and user traffic — makes it possible to offload traffic locally to edge data centers. In LTE networks this is impossible: all user traffic must be “dragged” into the network core, where the switching function is implemented (the analogue of UPF in 5G), and then sent back to the edge data center. That completely negates the main advantage of edge computing: low signal latency.

Despite five years of active 5G rollout around the world, even wealthy countries still face limited 5G coverage, especially stand-alone, or SA, networks. Only such networks allow all the advantages of 5G to be realized. 5G SA coverage remains concentrated mainly in large cities with the most affluent subscribers. And it seems this situation will persist for many years.

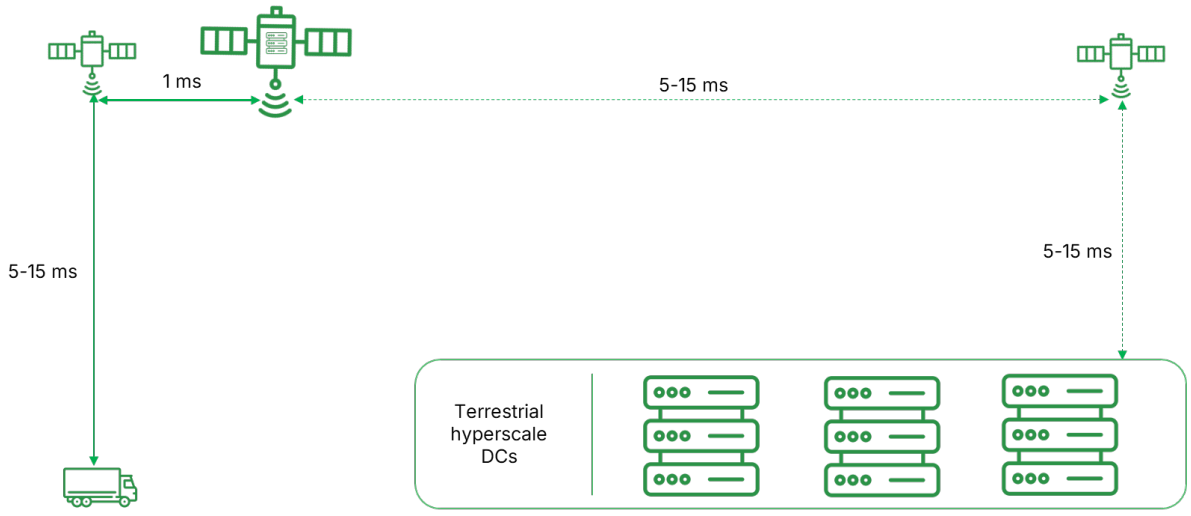
However, AI’s entry into the physical world, built above all on transport and logistics, already requires total broadband IP-network coverage of the Earth’s surface, including the oceans. If we leave aside truly exotic technical solutions such as giant airships in the stratosphere — development work on which was halted five years ago — only one option remains: low-Earth-orbit satellite constellations as the main network, with terrestrial 5G networks as reserve, localized coverage in regions with a high density of paying users.

Accordingly, in the case of satellites, edge computing must be located in the same place: in space, and also in low orbits. Again, from the point of view of latency: for terrestrial 5G networks, under certain conditions, it may theoretically be possible to achieve the real-time AI-control latency needed without edge computing; for satellite networks, this is impossible in principle.

But perhaps satellites are not suitable for such tasks at all? After all, even in relay mode, low orbits deliver round-trip delay of at least 20 ms. Yet we are talking about 10 ms and below, down to single milliseconds.

Here it is important to understand the architecture of the solution as a whole. As noted above, the computations most sensitive to latency must be performed on board. But not all of them — that would be economically inefficient. The AI agent itself must therefore be distributed: partly on board and partly in the edge data center. In such a configuration, latency between the onboard and edge compute nodes of up to 10 ms (up to 5 ms one way) is quite acceptable. Data sent to the terrestrial hyper-data center with end-to-end latency above 10 ms will be used only for LLM training; it will not be real-time operational data. For such cases, total latency of up to 50 ms or more is acceptable, as shown in Figure 1.

Figure 1. ICT system with space-based elements for AI control of transport



Relay mode, however, is indeed impossible for the task under discussion. To achieve latency of up to 10 ms, edge compute nodes must be in space as part of a communications-satellite constellation. Otherwise, the entire self-learning control system would have to be placed directly on the controlled object, making it autonomous not only in the sense that humans do not participate in control, but also in the sense that

intensive external information exchange is excluded. That would sharply limit the use of AI to very large, and therefore few, objects such as large ships — an unacceptable outcome.

Technically, adding compute satellites to low-orbit constellations of relay satellites poses no major problem. Since version 1.5, that is, since 2022, all new Starlink satellites have been equipped with optical, laser links to other satellites in the constellation, and can relay signals not only to Earth but also to another satellite, which may be a compute node rather than a relay. Yes, an edge-computing satellite must be larger than a relay satellite, but this is not a mega data center in space; it is an object roughly the size of Gagarin's Vostok spacecraft.

There is another major advantage to this configuration. With a single combined satellite constellation — relay satellites and compute satellites — one can avoid the most difficult and still unsolved problem of cross-domain orchestration: orchestration in a heterogeneous environment made up of domains owned by different economic actors. In this case, both space domains, as well as the domain of terrestrial hyperscalers, would be managed by one operating company.

Still, the technical problems of designing a compute satellite do exist, and they are quite critical. They are mainly connected with the need for innovative cooling systems and radiation protection for large computing equipment.

Technological Innovations for Compute Satellites

Clearly, a space data center is radically more expensive than a terrestrial one. There is the cost of the satellite, the cost of launching it into orbit, and the cost of operation. If the goal is to radically reduce the unit cost of computing, then putting a data center in space seems to move in exactly the opposite direction.

What is the current cost gap between computing on Earth and in space, and what possibilities exist for reducing it?

One gigawatt of computing power in orbit may cost $42.4 billion — almost three times more than on Earth — and a number of fundamental technical problems still have to be overcome. According to estimates, bringing the cost of orbital computing closer to that of terrestrial data centers would require launch costs to fall 18-fold, from the current $3,600/kg for Falcon 9 to $200/kg for Starship in the 2030s. At the same time, the launch-services market is not fully elastic: regulation is complex, the scale is limited, B2B and B2G customers have long sales cycles and wholesale contracts, and promising mega-constellations are vertically integrated. Up to 75% of Falcon 9 launches are for Starlink spacecraft, and Amazon Leo/Blue Origin and China's Guowang/Thousand Sails are moving along the same path. So the answer is yes, it may be possible — but not for everyone, only for vertically integrated giants.

The second necessary condition for bringing the cost of orbital data centers closer to terrestrial ones is a radical reduction in the cost of the compute spacecraft themselves. The cost of 1 kW of electricity for terrestrial data centers ranges from $570 to $3,000 per year, depending on local conditions. In orbit, yes, solar energy is free — but components still have to be purchased, the spacecraft developed and launched, and normal operation ensured. As a result, the cost of 1 kW on a Starlink spacecraft, according to calculations by senior Project Suncatcher staff, is $14,700 per year. A fivefold reduction is possible, taking into account mass production and an expected production learning curve for mass-produced spacecraft of 95% — a 5% reduction in unit production cost with every doubling of output. Overall, a multiple reduction, and even an order-of-magnitude fall in launch and spacecraft-production costs, is theoretically possible.

However, the extremely difficult task of heat removal in a vacuum remains. Solar panels in orbit are indeed 5-8 times more efficient than on the Earth's surface. But the absence of an atmosphere also prevents efficient heat removal. Preliminary calculations suggest that for every 20 kW of heat that must be dissipated, 20 square meters of radiating area are needed at a standard processor operating temperature of 80°C. Thus, the area of radiators alone for a 100 kW SpaceX AI Mini Sat could approach 100 square meters. This is technically feasible, though still ambitious. In his usual style, Musk called such a spacecraft "Mini", implying even larger systems in the future. For comparison, the electrical power of all solar arrays

on the ISS is 120 kW, with radiator area of 150 square meters. Musk's plans to develop customized radiation-resistant AI chips capable of operating at the higher temperature of 97°C are intended precisely to raise coolant temperature and overall radiator efficiency while reducing the dissipating area.

In addition to heat removal, there is also the need to fight radiation in order to avoid bit flips. The traditional approach of using radiation-hardened space-grade electronic components is unacceptable because of cost: a single RAD750 from BAE Systems, for example, costs about $200,000. Compute spacecraft for mega-constellations will need to use ordinary auto- or industrial-grade components, but with a protective shell. Google's laboratory tests of such coatings, with densities of 5-10 g/cm² in a particle accelerator, inspire optimism, but real flight tests still have to be carried out. That is why everyone is waiting keenly for the launch of two test spacecraft by Google and Planet under the Starcatcher project in 2027, with AI models deployed on board.

A More Expensive Element Does Not Necessarily Make the Whole System More Expensive

Even if all the innovations listed above are successfully brought to broad commercial use, a space data center will still always lose economically to a terrestrial one, even if we compare like with like — namely small edge data centers, terrestrial and space-based.

There is only one "but", and it completely changes how the problem looks: one must consider the economics of the entire system, not of a single element. As described above in Figure 1, by the system we mean at least two types of data centers — edge and root — as well as communication channels between the controlled object and the edge data center, and between the root data center and the edge data center. And we are not comparing hypothetical configurations of such a system, but realistic ones, taking into account bandwidth requirements, the latency budget, and the required level of system availability.

As noted above, for moving objects, which make the greatest contribution to the economics of digitalization, only two options meet the signal-latency requirements: 5G SA and low-orbit satellite communications such as Starlink. Both work only in a configuration with MEC. Given the coverage limitations of 5G SA, even on land the basic option for network connectivity can only be a satellite constellation. Stable 5G SA coverage with MEC in sparsely populated areas is a much greater fantasy than data centers in space. And air and water transport also remain; for them, satellites have no alternative at all.

Thus we have only one possible system configuration, and in that system there is one extremely expensive but necessary element: compute satellites. Paradoxically, the presence of this element creates opportunities to reduce the cost of the remaining parts of the system. These include much lower hardware requirements for the onboard control platform of an autonomous transport vehicle, provided stable connectivity exists — a critical factor given how many such controlled objects there will be. They also include savings in the bandwidth of communication channels, both radio and terrestrial wired, or optical, channels, and an almost complete absence of constraints on where root data centers can be located. There is no need to build data centers in energy-deficient regions, which can improve their economics severalfold.

Another factor, perhaps even more critical than the cost of compute satellites, is the efficiency of end-to-end orchestration in such a complex communication-and-computing system. As noted above, a configuration with space-based data centers does not complicate orchestration; on the contrary, it simplifies it. In this case, the need to use terrestrial broadband data networks is minimized or eliminated entirely, and all ICT-infrastructure elements can be deployed by a single economic actor under a unified technical policy.

In other words, the case may theoretically work, given the high unit economic effect of digitalization specifically in transport and logistics. Nor should we discount the role of personality in history: after all, there are people who, like Korolev, find it important not merely to live, but to live with passion.

Attempts at Implementation: SpaceX, xAI, TeraFab

In February 2026, SpaceX asked the U.S. regulator, the FCC, for permission to deploy a distributed orbital data center for AI workloads with total power of 100 GW, consisting of one million spacecraft. Each AI Mini Sat, weighing about one tonne, would have power of 100 kW — four to five times higher than the current Starlink V2 Mini generation. The satellites would be connected by Starlink laser links.

In theory, this would make it possible to connect both an enormous number of spacecraft and the existing terrestrial compute segment of the root data centers of the recently acquired xAI with the orbital segment. Largely for this compute mega-constellation, SpaceX/xAI and Tesla announced the launch of TeraFab: a vertically integrated factory for producing 2 nm GPUs with total capacity of 1 TW per year — that is, 1 billion AI chips. This includes processor design, lithography, fabrication, memory-chip production, packaging and testing. The first construction phase is estimated at $25 billion, and the first mass-produced chips are expected in 3-5 years. In the future, production is expected to scale not so much to cover needs on Earth — Tesla FSD/Dojo/robotaxi, Optimus, and xAI/Grok are forecast to account for only 20% — as to cover needs in space: 80% of production capacity for orbital data centers, and for the exploration of the Moon and Mars.

The merger of SpaceX and xAI, together with Musk's ambition for a million-satellite computing constellation — let us recall that Starlink currently has about 10,000 active spacecraft, which is already considered the world's largest orbital constellation — may be a PR move aimed at raising the market valuation of the combined company to $1.75 trillion ahead of the Starlink IPO. At the same time, the global production capacity of all leading AI-processor fabs does not exceed 20 GW per year — only 2% of the 1 TW Musk requires. These are extremely ambitious plans, even considering that Intel is expected to serve as the project's industrial partner, bringing a 1.8 nm process and experience in organizing mass contract manufacturing.

Nevertheless, amid the broad interest in AI, other tech giants are also taking a lively interest in orbital computing. Google, together with Planet Labs, has announced Project Suncatcher: an orbital cluster of 81 spacecraft based on Tensor Processing Units, or TPUs — specialized tensor processors designed to accelerate machine learning and neural-network-based computing. The first prototype spacecraft is planned for launch in 2027.

At the same time, Google invested in the startup Starcloud, which filed an application with the FCC to deploy a constellation of orbital data centers consisting of 88,000 spacecraft. One Starcloud-4 cluster in orbit, with 5 GW of power and based on NVIDIA Space-1 Vera Rubin compute modules, would measure 4 by 4 kilometers. The viability of such monolithic megastructures in low Earth orbit is hard to believe, given the current problem of space debris.

Jeff Bezos's Blue Origin is also planning an orbital compute mega-constellation, Project Sunrise, consisting of 51,600 spacecraft with full vertical integration. The platform for AI spacecraft will apparently be the heavy orbital tug Blue Ring; launches will use the New Glenn rocket; and data transmission to Earth will rely on laser links from the planned TeraWave constellation of 5,408 spacecraft, with total throughput of 6 Tbps.

China's Adaspace has also announced an AI mega-constellation of 2,800 spacecraft with computing capacities.

Clearly, some of the announced projects will be cancelled, and those that remain will not achieve their initially exaggerated goals. It could hardly be otherwise. But an objective need for relatively small compute satellites in space does exist — not for training LLMs, but for controlling physical objects with the help of AI agents — and it may well be met in the not-so-distant future.

A Final Word About the Future

AI supported by a distributed network-computing infrastructure with edge compute nodes in space is, in a very real sense, already Skynet. There is already a sufficient level of investment, a real need, and the technical capability to develop and deploy it.

We know from the film Terminator what is wrong with the Skynet story. So, returning to the beginning of this article, let us once again note that innovation in social and economic relations is no less important — indeed, it is more important — if technological breakthroughs are not to destroy humanity but to serve it. That, however, is a topic for a separate series of publications, one large enough to fill several issues of Connect magazine...

Source links from the original article

1. Andre Norton, Sargasso of Space: https://libcat.ru/knigi/fantastika-i-fjentezi/28809-andre-norton-sargassy-v-kosmose.html

2. N4A Analytics, autonomous systems: https://www.n4a.info/autonomous-systems-ru

3. SpaceX status: https://x.com/SpaceX/status/2046713419978453374

4. Cursor and SpaceX model training: https://cursor.com/blog/spacex-model-training

5. Connect magazine issues: https://www.connect-wit.ru/izdaniya-connect.html

6. Pro Arctic airship article: https://pro-arctic.ru/10/04/2015/expert/15168

7. Starlink latency map: https://starlink.com/map?view=latency

8. Autonomous vehicles and V2X latency note: 3-10 ms for autonomous vehicles, 5G standard for the V2X network layer.

9. Andrew McCalip, space data centers: https://andrewmccalip.com/space-datacenters

10. Project Suncatcher paper: https://arxiv.org/pdf/2511.19468

11. CBO learning curve reference: https://www.cbo.gov/publication/58794

12. Cooling and radiators reference: https://www.youtube.com/watch?v=FIQYU3m1e80

13. Luminix industry analysis: https://www.useluminix.com/reports/industry-analysis/data-centers-in-space

14. Google Research, space-based scalable AI infrastructure: https://research.google/blog/exploring-a-space-based-scalable-ai-infrastructure-system-design/

15. Yahoo Finance, SpaceX IPO/valuation: https://finance.yahoo.com/markets/stocks/articles/spacex-going-public-1-75-202500656.html

16. Yahoo Finance, SpaceX and xAI: https://finance.yahoo.com/sectors/technology/articles/heres-real-reason-spacex-teaming-131200500.html

17. Forbes, Intel and TeraFab: https://www.forbes.com/sites/jonmarkman/2026/04/10/intel-joins-terafab-to-build-elon-musks-25b-ai-chip-project/

18. Google Cloud TPU: https://cloud.google.com/tpu

19. Starcloud white paper: https://starcloudinc.github.io/wp.pdf

20. Habr, space debris: https://habr.com/ru/companies/ruvds/articles/595695/

21. Data Centre Magazine, Project Sunrise: https://datacentremagazine.com/news/project-sunrise-blue-origin-data-centre-space-race