Space Datacenters

Nathan Bleier and Rakesh Kumar

Abstract—Falling launch costs, growing satellite data volumes, and rising terrestrial power constraints have prompted serious investigation into moving datacenters to space. We look at the motivations, engineering challenges, architectural design space, and technology gaps that will determine whether space datacenters become a practical reality.

Index Terms—space datacenters, computing

For sixty years, computing has been an almost entirely terrestrial activity. Datacenters sit on the ground, draw power from the grid, and cool themselves with local air or water. That assumption is now being questioned. A confluence of factors — the explosive growth of space-based data generation, falling launch costs, latency constraints on deep-space missions, security vulnerabilities in satellite downlinks, and growing strain on terrestrial power grids — has prompted serious investigation into moving computation off-planet. Industry momentum is building rapidly: Google's Project Suncatcher aims to fly TPU-equipped satellite clusters, NVIDIA-backed Starcloud has already trained a large language model on an orbital GPU, and defense agencies are deploying constellations with onboard processing for missile tracking.

1 Why Build Space Datacenters?

Several converging trends are motivating the need for server-based computing in space (Fig. 1).

The cost to launch payloads to low Earth orbit has dropped dramatically over the past two decades. From 1970 to 2000, launch costs averaged $18,500 per kilogram, with large-scale satellite deployments via the Space Shuttle costing $54,500 per kilogram. Today, SpaceX's Falcon 9 has reduced this to $2,720 per kilogram, a twenty-fold reduction from historical prices. Falcon Heavy achieves $1,400 per kilogram.

This cost collapse has enabled an explosion in satellite deployments. In 2000, approximately 2,500 active satellites orbited Earth, operated by just 14 nations. By 2024, that number had grown to over 11,500 active satellites, with 91 new spacefaring nations entering the domain. This proliferation has created a downlink deficit: satellites generate far more data than they can transmit to Earth. A single high-resolution Earth observation satellite produces 1–2 terabytes of imagery per day, yet RF downlinks operating at 220 Mbit/s during brief ground passes can deliver only tens of gigabytes [1]. Optical ground links promise higher rates but still cannot close the gap at high resolutions, and suffer atmospheric attenuation from clouds, rain, and turbulence that limits their availability [2]. Applications drive the need for ever-finer resolution: precision agriculture requires 30 cm imagery to distinguish individual crop rows, while infrastructure monitoring and disaster response demand sub-meter detail to detect damage to roads and buildings [2].

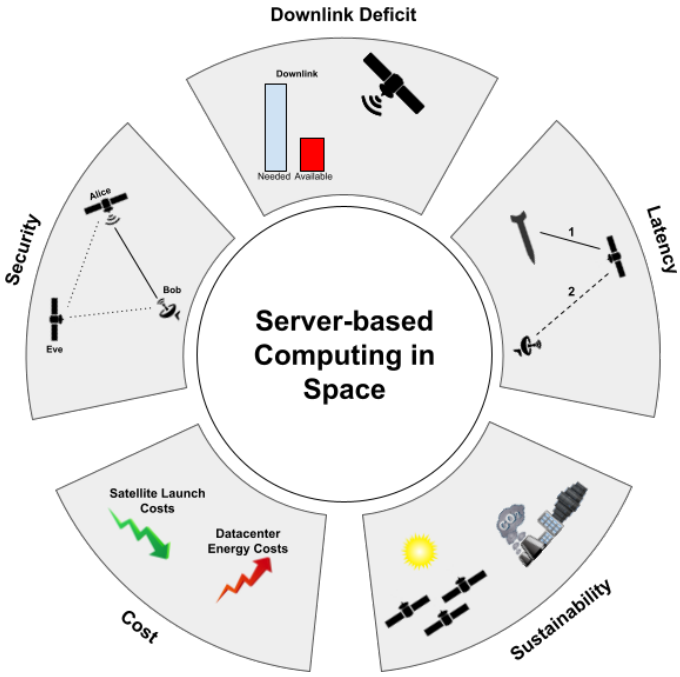


Fig. 1. Converging motivations for server-based computing in space: the growing downlink deficit, security vulnerabilities in satellite communications, latency constraints on deep-space missions, declining launch costs alongside rising datacenter energy costs, and sustainability concerns with terrestrial power generation. From [1].

The scale of the problem is immense: a global constellation imaging at 10 cm resolution every 10 minutes would require over 9 million simultaneous 220 Mbit/s channels (Fig. 4). Even with 95% early discard, high-resolution missions at 30 cm or finer face downlink deficits approaching 100%: nearly all generated data must be discarded due to insufficient capacity (Fig. 3) [2]. Moving computation to space to filter, clean, or process data directly mitigates this bottleneck by transmitting only results rather than raw data (Fig. 2).

Security provides another compelling motivation. Satel-

- *N. Bleier is with the Department of Computer Science and Engineering, University of Michigan.*
- *R. Kumar is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign.*

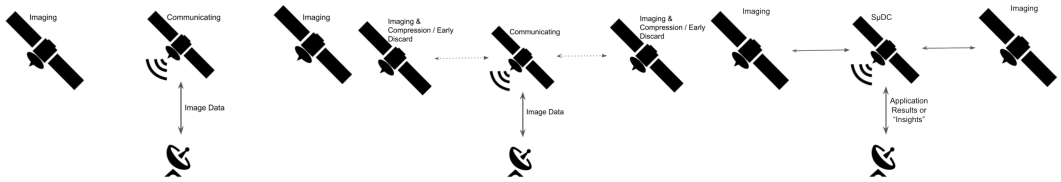
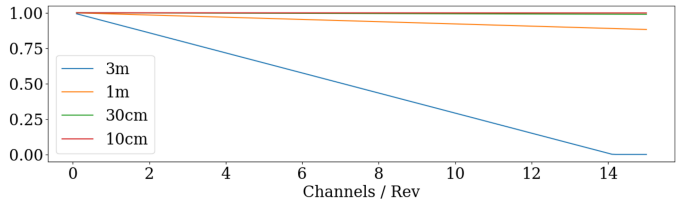
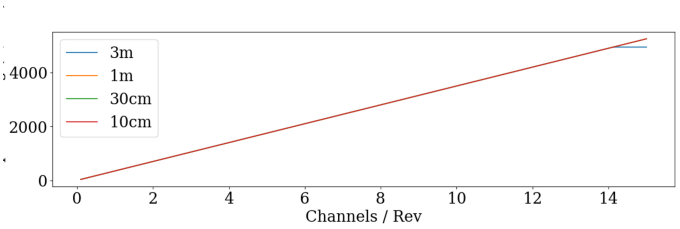


Fig. 2. Three approaches to satellite data handling: (a) a conventional Earth observation constellation downlinks all raw imagery, (b) compression and early discard reduce data volume but remain insufficient at high resolutions, and (c) a space microdatacenter (SµDC) processes data in-space and downlinks only application results. From [2].



(a) Downlink Deficit (DD)



(b) Time spent downlinking (per satellite per revolution)

Fig. 3. Downlink deficit (top) and time spent downlinking per satellite per revolution (bottom) at different spatial resolutions, assuming a 220 Mbit/s channel and 95% early discard. At resolutions finer than 1 m, nearly all data must be discarded regardless of available downlink channels. From [2].

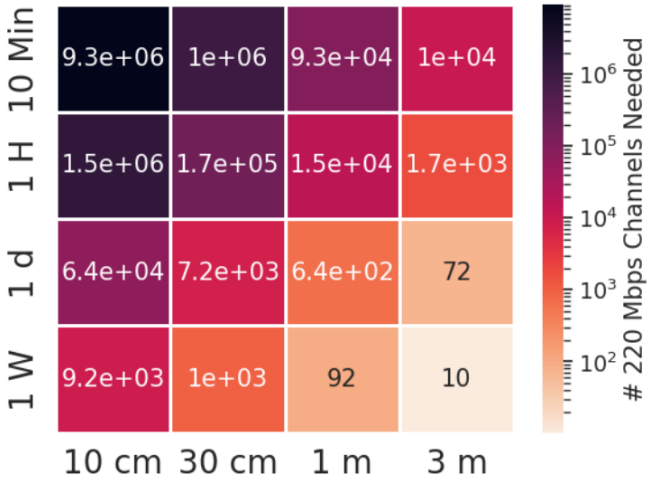


Fig. 4. Number of 220 Mbit/s downlink channels required for a global Earth observation constellation at varying spatial resolutions and revisit cadences. High-resolution, high-frequency monitoring demands millions of channels, far exceeding any feasible ground network. From [2].

lite communications, even when encrypted, leak information through side channels. By processing data in space, only insights rather than raw observations need be downlinked, eliminating the RF emissions that adversaries can intercept and analyze.

Another motivation is latency. Not all spacecraft are close to home, causing relatively large delays for communicating with Earth, limiting what they can do. For example, process control for any in-space manufacturing cannot tolerate round-trip delays to ground stations. Hypersonic missiles traveling through space must be detected and counteractions taken at latencies lower than the time it takes to communicate with Earth. NASA's Mars rovers cannot be driven in real-time and must instead rely on autonomous hazard avoidance software to navigate between waypoints. Compute co-located with a Mars mission, in orbit around Mars, would eliminate the Earth roundtrip entirely, enabling closed-loop control at speeds that ground-in-the-loop operation cannot achieve. In-space computing allows support for applications with low latency requirements.

Space also offers abundant energy. By placing energy production in space, we can guarantee 24/7 access to solar light, unimpeded by atmosphere or the umbral shadow of Earth. Solar panels in orbit produce approximately eight times as much electricity as their terrestrial equivalents, and the vacuum of space serves as an infinite heat sink for thermal management, reducing the need for water-intensive cooling towers [3].

Finally, space computing offers sustainability advantages. While embodied carbon from manufacturing and launch remains an issue, operational carbon of in-space computing is zero.

2 What Makes Building Space Datacenters Challenging?

Building computing infrastructure in space faces several challenges. The launch and placement process subject payloads to extreme noise, acceleration, and transient shocks, creating a high risk of failure for any computing and other equipment. Once deployed, the computing infrastructure operates in an environment where conventional maintenance is impossible. As one DARPA program manager observed, "There is no other area of human activity where we build something that's worth a half-billion dollars or a billion dollars, and never look at it again, never fix it, and never upgrade it." This reality demands extraordinary upfront reliability, yet the space environment actively degrades systems over time, due to drag and collisions with space debris, the population of which is growing rapidly. Radiation presents an additional challenge: high-energy particles cause both cumulative damage (total

ionizing dose) and transient errors (single-event upsets) in electronics. Fortunately, commercial CMOS logic processes have grown more radiation-tolerant with scaling, and the industry has largely moved away from expensive radiation-hardened components toward commercial-off-the-shelf hardware with system-level fault tolerance [1]. SpaceX's Dragon capsule carries 54 processors across six flight computers, all commercial x86 parts running Linux, using triple-redundant voting to detect and recover from radiation-induced errors. Hardware-level error detection and correction remains important for memory, which unlike logic does not benefit from electrical, logical, or timing masking of transient upsets.

Communication in space differs fundamentally from terrestrial datacenters, where servers connect through high-bandwidth, low-latency wired networks. Orbital systems must rely entirely on wireless links, either radio frequency or free-space optical, that introduce propagation delays, pointing requirements, and susceptibility to interference. Ground station contacts are fleeting: satellites in low Earth orbit typically have communication windows lasting only minutes per pass, with the physical configuration of orbits, ground station locations, and antenna capabilities constraining when and how much data can be exchanged [4]. Inter-satellite links (ISLs) enable direct communication between spacecraft without routing through ground stations, but maintaining alignment between rapidly moving platforms presents significant engineering challenges.

Space datacenters may also face a temporal continuity challenge that terrestrial systems do not. A ground-based datacenter runs continuously, maintains state indefinitely, and is always reachable. Space-based systems spread over multiple planes or altitudes may move in and out of mutual line-of-sight, requiring, in some cases, state hand-offs between spacecraft as the network topology shifts with each orbital period. Ground station contacts are similarly fleeting. Designing software and scheduling frameworks for such systems is a challenge with few terrestrial precedents.

3 A Taxonomy of Space Datacenter Architectures

Space datacenter proposals can be broadly classified along three dimensions: scale, topology, and physical architecture.

Scale refers to the total compute capacity of the system. At one end, space microdatacenters embed modest compute on individual satellites in the 500-watt to 10-kilowatt power range, processing data generated by the remote sensing constellations they support [2]. Even at these scales, compute hardware represents only 1–5% of total system cost, with solar arrays, thermal management, and structural components dominating the budget [1]. At the other end, mega datacenter concepts like Starcloud's 5-gigawatt vision or the EU's ASCEND study targeting 1 GW by 2050 envision facilities comparable to the largest terrestrial hyperscalers [3], [5]. The International Space Station, at 420 metric tons and 100 kilowatts, is the largest structure humans have built in orbit; a gigawatt-scale datacenter would dwarf it.

Topology describes how logical computers are organized. A *centralized* system is a single logical computer that performs one computing task at a time. This ranges from a single server to an entire datacenter operating as one computer on a single task, such as AI training. A *constellation* is a distributed system of multiple logical computers that are linked but capable of performing tasks independently and in concert, analogous to a networked cluster of servers. China's Three-Body constellation, with 2,800 satellites each carrying an 8-billion-parameter model for on-board inference, exemplifies the constellation approach: each node operates autonomously [6]. SpaceX's proposed million-satellite orbital data center would also adopt a constellation topology, with nodes linked by laser inter-satellite connections for distributed AI training [7].

Physical architecture describes how each logical computer is implemented, independent of scale. A *monolithic* system integrates all resources (compute, memory, power, cooling) into a single physical platform. This applies at any scale: a space microdatacenter contained within a single satellite is monolithic, and so is a gigawatt-scale platform like Starcloud that places GPU servers on one structure with wired NVLink interconnects [3]. A *disaggregated* system composes each logical computer by pooling resources from multiple separate physical spacecraft, connected by high-bandwidth links. Google's Project Suncatcher exemplifies disaggregation: each logical compute unit is an 81-satellite cluster operating 100–200 meters apart, connected by free-space optical links [8]. Disaggregation improves resiliency (a single satellite failure reduces capacity rather than causing total loss), accelerates technology refresh through shorter development cycles, and becomes increasingly attractive as launch costs fall. The cost is higher-latency interconnects and higher parasitic mass, since there is limited sharing of structure, power, and thermal systems across the compute elements.

These three dimensions (Fig. 5) are largely orthogonal. A monolithic microdatacenter on a single satellite occupies a very different point in the design space than a disaggregated constellation of clustered spacecraft. Existing proposals occupy only a few points within this space, and the architectural tradeoffs at each point remain largely unexplored.

4 Making Space Datacenters a Reality

The economic case for orbital computing depends heavily on the application. Some use cases already make sense at today's launch costs: co-flying compute satellites for space telescopes, onboard processing for Earth observation constellations, and edge computing for latency-sensitive defense applications. These scenarios derive value from proximity to space-based sensors or instruments, not from competing with terrestrial datacenters on cost-per-flop [1], [9]. General-purpose cloud computing in orbit faces a higher bar. Google's Project Suncatcher analysis suggests space-based AI training becomes competitive when launch costs drop below $200 per kilogram, roughly an order of magnitude below current Falcon 9 prices [10]. SpaceX's Starship promises such reductions through full reusability, but the system remains in development. Launch availability also matters: a mature space datacenter industry would require launches to become routine, measured in thousands of flights per year rather than the roughly one hundred that occur today.

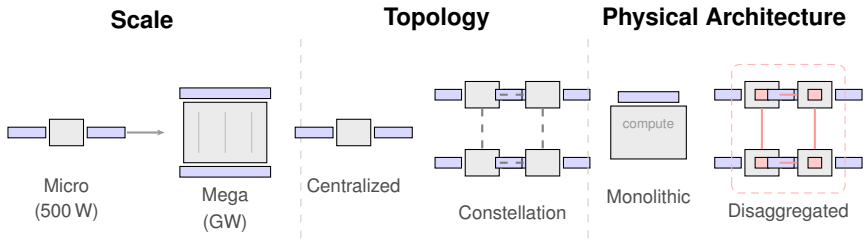


Fig. 5. Three dimensions for classifying space datacenter architectures. Scale ranges from single-satellite microdatacenters to gigawatt-class megadatacenters. Topology distinguishes centralized systems (one logical computer) from constellations (multiple cooperating nodes). Physical architecture distinguishes monolithic platforms from disaggregated systems where the computer is composed from multiple spacecraft (red).

Technology gaps span compute, memory, and communication. The space cost structure inverts traditional semiconductor economics (power costs dominate over silicon in space-based computing) yet the industry has historically optimized for the opposite: performance-per-dollar rather than performance-per-watt. However, terrestrial datacenters are increasingly facing similar constraints. Hyperscale operators are no longer limited by capital expenditure but by power availability: grid capacity, power purchase agreements, and generator infrastructure now determine where and how large datacenters can be built. This convergence suggests that investment in power-efficient architectures will benefit both terrestrial and orbital computing, potentially accelerating progress on space-suitable designs. Memory presents an acute challenge: HBM provides the bandwidth that modern workloads demand but cannot tolerate the radiation environment, while radiation-tolerant alternatives like MRAM and ReRAM lack the speed and energy efficiency of SRAM. Custom SRAM-based architectures offer a path forward but require significant investment in design and fabrication [9]. Inter-satellite communication has progressed rapidly (Starlink's 100 Gbps optical links demonstrate what is possible) but the gap between demonstrated point-to-point links and the terabit-scale switched fabrics of terrestrial datacenters remains vast. Delay-tolerant networking protocols, developed over decades for interplanetary communication, offer architectural principles for systems that must operate despite intermittent connectivity and variable latency; extending these concepts to support datacenter-class workloads in orbit remains an open research challenge.

Reliability expectations must evolve in both directions. Traditional space systems are designed for decade-long missions without maintenance, driving extreme conservatism in component selection and massive upfront qualification costs. Commercial computing hardware, by contrast, assumes failures will be handled through replacement, an option unavailable in orbit. Neither model fits space datacenters well. Disaggregated architectures offer a middle path: by distributing computation across many modest satellites rather than concentrating it in a few exquisite platforms, systems can tolerate individual failures gracefully. If one node fails, capacity degrades incrementally rather than catastrophically. This approach also enables technology refresh: older satellites can be decommissioned and replaced with newer hardware on a rolling basis, rather than locking an entire constellation into decade-old technology. On-orbit servicing and debris removal remain nascent capabilities, but as they mature, the economics of space computing will shift further toward replaceable, upgradeable architectures.

The space industry's manufacturing base must scale by orders of magnitude. Traditional aerospace operates in a low-volume, high-touch paradigm: satellites are hand-built by specialized technicians, individually tested for months, and launched one or two at a time. Modern constellations have begun to change this: Starlink, Planet, Iridium, and others now produce satellites on assembly lines and launch dozens at once. But even their manufacturing throughput would be modest by terrestrial electronics standards. A space datacenter industry deploying tens of thousands of compute nodes would require fabrication, integration, and test infrastructure that does not yet exist. The supply chain poses similar challenges: radiation-tolerant components are produced in small quantities at high prices, and many critical parts have single-source suppliers. Scaling production would reduce per-unit costs, but achieving that scale requires demand that cannot materialize until costs fall, a classic chicken- and-egg problem that only sustained investment or breakthrough applications can resolve.

Political dynamics will also shape the trajectory. Space infrastructure has historically triggered competitive responses among nations, and orbital datacenters are unlikely to be an exception. If one country establishes significant compute capacity in orbit, others will face pressure to match it, both for strategic autonomy and to avoid dependence on a rival's infrastructure. Export controls, spectrum allocation, and orbital debris governance add further complexity.

Finally, the problems worth solving in space must become clearer. Today's compelling use cases share a common trait: they process data that originates in orbit. Earth observation satellites filter terabytes of imagery to extract gigabytes of insight. Space telescopes require real-time wavefront control that ground links cannot support due to latency. Defense systems demand continuous processing that intermittent ground station contacts cannot provide, and resilience against adversaries who might target ground infrastructure. These applications justify orbital computing not by competing with terrestrial datacenters on cost, but by enabling capabilities that would otherwise be impossible. The vision of general-purpose cloud computing in space, training AI models on orbital GPU farms or running enterprise workloads above the atmosphere, remains speculative. It requires not only the cost reductions and technology advances discussed above, but also workloads for which space offers genuine advantages beyond abundant solar power. Whether such workloads emerge, or whether orbital computing remains confined to processing space-generated

data, will ultimately determine the scale and shape of the industry.

REFERENCES

[1] N. Bleier, R. Eason, M. Lembeck, and R. Kumar, "Architecting space microdatacenters: A system-level approach," in *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2025.

[2] N. Bleier, M. H. Mubarik, G. R. Swenson, and R. Kumar, "Space microdatacenters," in *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 900–915.

[3] NVIDIA, "How Starcloud is bringing data centers to outer space," NVIDIA Blog, 2025.

[4] B. Denby and B. Lucia, "Orbital edge computing: Nanosatellite constellations as a new class of computer system," in *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 2020, best Paper Award.

[5] Thales Alenia Space, "Thales alenia space reveals results of ASCEND feasibility study on space data centers," Thales Alenia Space Press Release, 2024.

[6] A. Jones, "China launches first of 2,800 satellites for AI space computing constellation," SpaceNews, 2025.

[7] J. Foust, "SpaceX files plans for million-satellite orbital data center constellation," SpaceNews, 2026.

[8] Google, "Meet project suncatcher, a research moonshot to scale machine learning compute in space," Google Blog, 2025.

[9] B. Lyu, V. Manjarekar, and N. Bleier, "Compute system organization for high frequency high order wavefront sensing and control," *Journal of Astronomical Telescopes, Instruments, and Systems*, 2026, submitted.

[10] Google Research, "Exploring a space-based, scalable AI infrastructure system design," Google, Tech. Rep., 2025, project Suncatcher Technical Report.