2025 IEEE International Symposium on High-Performance Computer Architecture

Architecting Space Microdatacenters: A System-level Approach

Nathan Bleier†
Department of Electrical Engineering & Computer Science
University of Michigan
Ann Arbor, MI

Rick Eason
Department of Aerospace Engineering
University of Illinois
Urbana, IL

Michael Lembeck
Department of Aerospace Engineering
University of Illinois
Urbana, IL

Rakesh Kumar
Department of Electrical and Computer Engineering
University of Illinois
Urbana, IL

***Abstract*—Server-based computing in space has been recently proposed due to potential benefits in terms of capability, latency, security, sustainability, and cost. Despite this, there has been no work asking the question: how should we architect systems for server-based computing in space when considering overall cost. This paper presents a Total Cost of Ownership (TCO)-based approach to architecture of server-based computing systems for space (Space Microdatacenters - SµDC) for processing data produced by low Earth orbit (LEO)-based Earth observation (EO) satellites. We show that power of compute is the primary factor in determining SµDC TCO, though the dependence is sublinear. Second, the impact of compute mass, monetary cost, and communication on TCO is relatively insignificant. Third, architectures with the highest $\frac{\text{FLOPs}}{\text{W}}$ provide much higher performance per TCO $ even if they have poor $\frac{\text{FLOPs}}{\$}$. We leverage these insights to advocate *extreme heterogeneity* designs for SµDCs. These designs reduce SµDC TCO by $116\times$ in spite of poor $\frac{\text{FLOPs}}{\$}$ characteristics. We also show that (a) *collaborative compute constellations* — constellations in which EO satellites are also equipped with compute hardware — further improve SµDC TCO by 1.31 to $1.74\times$, (b) a *distributed architecture* reduces TCO by 10% over a monolithic architecture, and (c) low monetary cost of compute can be leveraged to provide *near zero cost compute overprovisioning* which improves an SµDC's availability significantly and supports graceful degradation. Overall, this is the first paper on cost-aware architecture and optimization of a SµDC.**

I. Introduction

Space is increasingly being suggested as a new frontier for server-based computing [9], [35], [46], [53], [66] (Figure 1). First, there are considerable bandwidth and cost bottlenecks to moving large amounts of satellite generated data to Earth for server-based processing, instead of processing it in space itself on dedicated compute servers [9], [19]. Second, moving satellite-generated data to Earth before processing increases latency - limiting the use of space data for critical low-latency applications such as aircraft detection [33], [69], realtime traffic monitoring [39], flood and forest fire detection [85], etc. Third, moving space data to Earth for processing has security implications - providing a sophisticated adversary opportunities to disrupt, decipher, or contaminate communication and processing [96]. Processing satellite data on space-based servers limits security vulnerabilities [99]. There are also arguments that space-based computing would be more sustainable [50], [66] even for terrestrial data since it would be powered primarily by the sun. It may also be cheaper as the launch costs decrease [37].

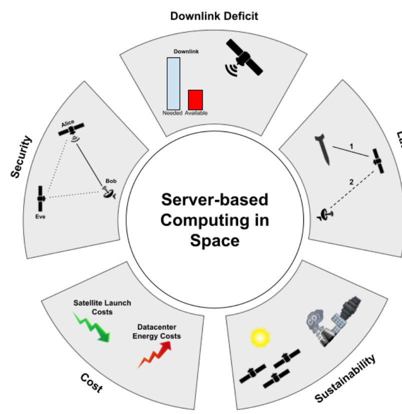


Figure 1: Server-based computing in space: key motivations.

Despite this, there has been no work asking the question: how should one architect systems for server-based computing in space when considering overall cost? It is unclear what the right metrics are, what the key design parameters are, and what the new opportunities and challenges are, especially with respect to terrestrial datacenters.

† The author's affiliation at time of research was the *Department of Electrical and Computer Engineering* at University of Illinois in Urbana, IL.

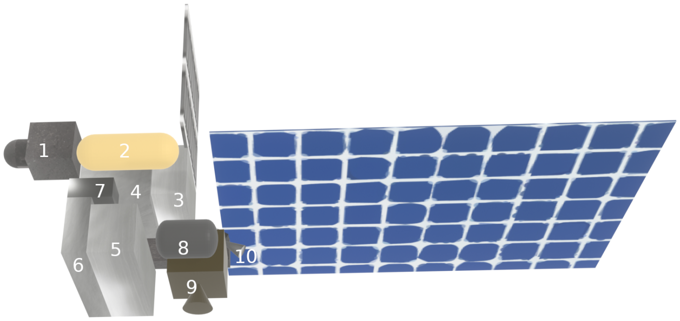


Figure 2: A 4 kW SμDC (w/o bus frame and wiring). In addition to the solar array (right) and radiator (top), The components are 1) ISL Transceiver, 2) N_2 tank, 3) heat exchange system and heat pump, 4) payload compute vessel, 5) reaction wheels and ADCS, 6) battery and PDU, 7) CDH and TT&C, 8) fuel tank, 9) thruster, 10) star tracker.

This paper presents a system-level approach to cost-aware architecture of server-based computing systems for space for processing data produced by low Earth orbit (LEO)-based Earth observation (EO) satellites. We assume that servers in space will initially be housed inside a self-contained satellite that supports power generation and distribution, cooling, etc., and, therefore, resembles a terrestrial microdatacenter in its organization [1]. A recent work making a case for SμDCs has also considered the same organization for servers in space [9]. Figure 2 shows a representative 4 kW SμDC.

To identify the key metric for SμDC design, as well as opportunities and challenges for architecture and optimization, we rely on a TCO (total cost of ownership)-driven approach. TCO analysis allows one to estimate the overall costs of acquiring and operating a datacenter [30], as well as impact of different metrics and design components on the TCO. We extended an existing free satellite cost model to include both the cost of compute hardware itself and the substantial cost of required support systems, such as power generation, thermal management, and free space optics (FSO) based inter-satellite-links (ISLs).

With this TCO model, we present the first TCO analysis for a SμDC. Our TCO analysis produces a number of key insights. First, the analysis shows that the cost of a SμDC depends strongly but sublinearly on the power of compute (and, therefore, computational capability). Second, the dependence on other factors such as compute mass, monetary cost, and communication on TCO is relatively insignificant - this is different from terrestrial datacenters where monetary cost of computation as well communication costs are a significant fraction of the TCO [30]. Third, we find that the architectures with the highest $\frac{\text{FLOPs}}{\text{W}}$ provide significant improvements in performance per TCO dollar even if these architectures have poor $\frac{\text{FLOPs}}{\$}$, and thus special purpose, accelerator-based architectures may be particularly attractive in a SμDC compared to more general purpose computational architectures.

We leverage these insights to perform a design space exploration of accelerator-based architectures to optimize TCO across a selection of key Earth observation tasks (i.e., image classification, object recognition, image regression, and image segmentation). We show that extremely heterogeneous designs consisting of multiple accelerators, each targeting individual neural network layers, minimizes TCO — a major departure from a terrestrial datacenter. These designs reduce SμDC TCO by 116× in spite of poor $\frac{\text{FLOPs}}{\$}$ characteristics.

We also present three system architecture optimizations that leverage the insights our TCO anaysis generates. First, we show that collaborative compute constellations — constellations in which EO satellites are also equipped with compute hardware to perform some initial computation on data before offloading data to an SμDC — further improve SμDC TCO by 1.31 to 1.74×. Second, we show that a distributed architecture with several small SμDCs may minimize TCO vs a monolithic architecture with one large SμDC supporting the same aggregate computational capability due to experience effects. A distributed architecture has an added advantage of supporting graceful degradation in case of failures. Finally, we show that compute over-provisioning can be supported at near zero cost on a SμDC due to low monetary cost of compute. Among other things, this can be used to improve the availability of a SμDC through sparing, for example.

Overall, this is the first paper on cost-aware architecture and optimization of a SμDC and contributes the first SμDC TCO model, the first SμDC TCO analysis, several key insights to drive SμDC architecture, and several SμDC architectural optimizations, such as extremely heterogeneous design, collaborative compute constellation distributed SμDC architecture, and near zero cost overprovisioning - with a quantification of their benefits.

II. Modeling TCO of a Space Microdatacenter

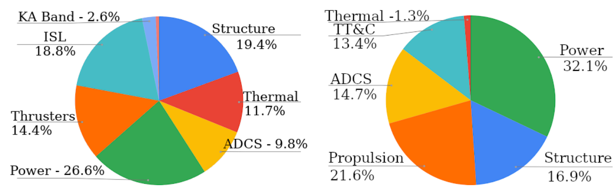
To model TCO of a SμDC, we build on the Small Satellite Cost Model (SSCM) [4]. SSCM is a parametric, cost-estimating relationship (CER) based model built and maintained by the Aerospace Corporation that is used to estimate the total cost of small satellites (i.e., satellites < 1000 kg). The model is based on a survey of 68 small satellites that captures satellite design parameters and satellite subsystem and operational costs. This data serves as the input to nonlinear regression models to identify the CER parameters. Although the survey looks at a large number of satellite design parameters (> 200), the regression models eliminate all but seventeen of the parameters, based on statistical analysis and experience.

The SSCM CERs estimate costs of all satellite subsystems and engineering processes except payload, which SSCM does not attempt to estimate. Each CER is further split into two cost categories: non-recurring (NRE) and recurring (RE) costs. NRE cost is associated with the design, verification, test, and management costs associated with producing a satellite prototype, as well as design and procurement of ground station equipment, while RE costs are the costs associated with procurement, deployment (i.e., launch costs), and lifetime management (including decomissioning) of the satellite itself. That is to say, the total cost (modulo payload) of the first satellite is equal to the sum of the NRE and RE costs of each CER, while the total cost of each subsequent satellite is given by RE costs alone.

SSCM is a staple tool of satellite designers, including those from NASA and MIT for cost modeling of Earth observation satellite constellations [59], by ESA for cost estimations for commercial launch vehicle [22], by the Naval Academy Small Satellite Program [40], by Satrec Initiative and NARA Space Technology [43], by Ienai Space [91], and by the AeroSpace Corpoation [45]. In spite of its popularity among aerospace industry, space agencies, military, and academia, SSCM, by itself, is inadequate for our needs. It does not model data processing subsystem or FSO, or their impact on other satellite subsystems. We extended SSCM to create a TCO model for SµDCs. The model differs from SSCM primarily in modeling of power and mass costs of computation and FSO-based communication, and impact of power and mass of computation and communication on other subsystems such as power generation, cooling, etc. Table I lists how values for different parameters in a SµDC cost model are determined. Below we provide details for some of the key derivations. The reader may skip directly to Section III without loss of continuity.

As mentioned above, we extend SSCM to include the cost of data processing and FSO. Our extensions were developed in consultation with five outside satellite design experts, including an original contributor to SSCM, and key mission designers for numerous small, Earth observation satellite missions, and for large satellite missions, including NASA's Galileo Jupiter Orbiter and the Wake Shield Facility. Our model increases the required power generation capacity of the satellite (SµDCs, being LEO-based, are solar powered; distant missions may use nuclear batteries [63]) as well as its cooling requirement (we assume radiator-based cooling - Section III-B) by the power cost of computation. Heat pump power (to pump heat to the radiator to increase its temperature - Section III-B) is determined based on the heat pump's Coefficient of Performance (CoP), which, in turn, is determined by radiator and ambient temperatures. Beginning of Life (BOL) system power, the amount of power generation that must be supported, is determined by End of Life (EOL) system power - the power requirements at the end of the satellite lifetime, the solar cell technology, and the orbit specific solar panel efficiency decay rate (generally $\leq 3\%$ annual loss) - to account for the fact that efficiency of solar panels degrades over time, and the satellite's projected lifetime.

Increased compute power also increases the mass and power of other subsystems. We model relationships between additional compute power and other model inputs: power subsystem mass, structural subsystem mass, Attitude Determination and Control System (ADCS) mass, propulsion system mass, fuel mass, command and data handling system mass, and thermal subsystem mass. Fuel mass is calculated using the rocket equation [56] — $m_{fuel} = m_{dry}\left(1 + e^{\Delta v/v_e}\right)$ where Δv is a mission dependent value, v_e is the exhaust velocity of the rocket engine, and m_{dry} is the 'dry' mass of the satellite (i.e., total satellite mass excluding fuel). The direct monetary cost of compute was added as well. To model the cost of FSO communication, we add FSO mass and power requirements to the mass and power of the Command and Data Handling (C&DH) subsystem. As C&DH cost estimates grow with communication data rates, and since FSO data rates are orders of magnitude better than RF, and the SSCM assumes RF communications, we first downscale the FSO data rate by the bandwidth ratio between FSO and X-band RF communications. Failure to do this downscaling results in unreasonably high C&DH cost estimates. Optical ISLs mass, power, and data rates are based on published values for *existing* commercial systems for LEO-LEO and LEO-GEO/MEO (medium Earth orbit) communication.



(a) SEER-Space (b) SSCM-SµDC

Figure 3: Subsystem cost breakdowns.

We compare the model against results produced by a commercial satellite cost model. Figure 3 shows the subsystem cost breakdown of a 4 kW SµDC as estimated by our SµDC variant of SSCM, as well as SEER-Space, a commercial spacecraft cost estimation model [27]. At first glance, the charts may appear quite different, but these differences are, in fact, minor or due to accounting differences. For example, in the SEER-Space model, the thermal subsystem is selected as 'active', rather than 'passive', as the designed SµDC uses an actively powered heat pump. However, in the SEER regression data set, active cooling was rare, which means the data are regressed against mostly passive cooling systems. Instead, in the SSCM-SµDC model, the power cost of active thermal management is included as a cost of the power subsystem. This means that SEER-Space distributes the cost more evenly between the thermal and power subsystems, while SSCM-SµDC concentrates the cost in the power subsystem. However, the sum of these two subsystems makes up 34.3% and 33.4% — a percent difference of less than 3%.

Structure subsystem costs are also fairly equal — they have a 7.4% percent difference. SEER-Space underestimates the ADCS cost relative to SSCM-SµDC. This is because SSCM-SµDC enables fine-grained control over ADCS performance parameters, which allows specifying a SµDC to have 50 microminute-of-angle pointing capabilities, while such fine-tuned control is not possible in stock SEER-Space. Similarly, SSCM-SµDC overestimates the cost of propulsion relative to SEER-Space. This is due to the fact that SSCM-SµDC is designed around conventional monopropellant and bipropellant chemical thrusters, rather than ion thrusters, while SEER-Space is parameterized to accept ion thrusters. Thus, SSCM-SµDC overestimates the cost of propulsion for larger, high power-generation satellites such as a 4 kW SµDC.

SSCM-SuDC results also match well with the costs we have observed in our own previous and ongoing satellite design and launch efforts in the Laboratory for Advanced Space Systems at Illinois (LASSI).

We will provide our model, implemented as a Rust library, on request to anyone who can present an SSCM license. Public access to SSCM-SuDC will lead to further community-driven validation.

III. TCO analysis of a Space Microdatacenter

Using the TCO model, we first study how the mission cost of a space microdatacenter changes based on the power budget devoted to compute (and, therefore, computational capability). Figure 4 shows how TCO increases

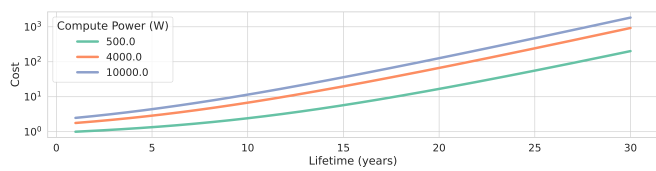


Figure 4: TCO vs Lifetime for 500 W, 4 kW, and 10 kW SµDCs relative to the 500 W SµDC with a one year lifetime.

with SµDC lifetime. For long lifetime missions, the cost grows superlinearly. The superlinear cost increase is driven by several factors. First, NRE and RE costs increase with lifetime, as additional reliability features are required. Second, fuel mass needed for station-keeping increases linearly with lifetime. Third, BOL power generation requirements increase exponentially, which increases power subsystem mass and, in turn, ADCS and fuel mass. For the rest of the analysis, we use five year lifetimes as it corresponds to roughly two technology generations for compute[1], while also limiting the total ionizing dose received by the non-radiation hardened computers. For similar reasons, Starlink targets a five year lifetime for its satellites.

Figure 5 shows the total and subsystem level costs of SµDCs from 0.5 kW to 10 kW. The costs are normalized against the total cost of a 0.5 kW SµDC. The results show that the **power of compute is the primary factor in determining TCO of a SµDC**. Cost can increase by over 3× when the compute power is changed from 0.5 kW to 10 kW. It is interesting to note is that the dependence is sublinear. A 20× increase in power corresponds with $< 4\times$ increase in total cost. There are two key reasons for this sublinearity. First, several technologies used in satellites have stabilized (e.g., solar panels, thermal systems, etc). So, increase in compute power does not increase their cost much — e.g., costs associated with design, test, and integration of these subsystems scale sublinearly. Second, total satellite mass, which affects both launch costs and design of different subsystems and therefore TCO, scales only slowly with compute power, as some satellite components require minimal scaling (e.g., C&DH, TT&C, ADCS, etc) while other require only limited amounts of scaling (e.g., propulsion fuel mass must scale, but only proportionally with the increase in overall mass — i.e., Amdahl's law is in effect). This means that non-compute costs increase sublinearly as the size of the compute payload increases.

Figure 5 also shows that, unlike terrestrial datacenters, where hardware costs are a majority portion of TCO [30], **the impact of monetary cost of compute on SµDC TCO is relatively insignificant**. Mass produced, commercial-off-the-shelf hardware, such as commodity NVIDIA GPU servers, have very low cost compared to the custom components common on satellites. As a result, the computer hardware cost of a SµDC is < 1% of TCO. Further, computer hardware is light — making up only a few percentage of total mass (Figure 6). Even after packaging, PCB integration, adding cooling, etc., an NVIDIA A40 GPU server has specific power of $> 35\,\frac{\text{W}}{\text{kg}}$. Further, the specific power of computer chips is very high: a 300 mm wafer has mass of ∼ 125 g but represents tens of kW compute power. This means adding additional, redundant *chips* to a system has negligible impact on both TCO and satellite mass (which itself affects TCO).

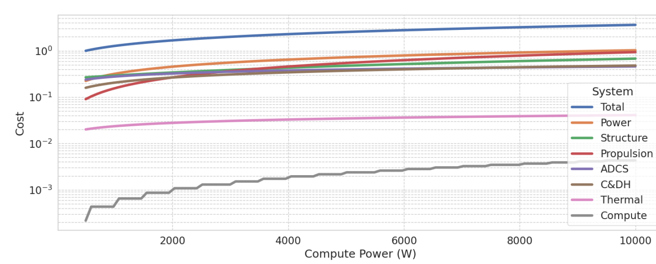


Figure 5: TCO vs Compute Power. Costs are relative to the total cost of a 500 W SµDC.

Figure 7 shows the relationship between TCO and ISL data rates assuming today's FSO power efficiencies [9]. For context, Figure 8 shows the ISL channel capacities needed to saturate various levels of compute hardware, based on application profiling of representative space data-based applications on RTX 3090 GPUs [9] (similar to many previous works [9], [18], [19], we assume that computation will be performed on GPUs - however, all our analyses and

[1]Five years is also approximately the time of two generations of NVIDIA's flagship neural network accelerator GPU line (e.g., A100 to B100)

TABLE I: Derivations for SSCM-SµDC Input Parameters.

Parameter	Equation	Paramter	Equation
Compute Power Consumption (W)	P_{comp}	Compute Ambient Temperature (K)	T_{amb}
Radiator Temperature (K)	T_{rad}	ISL Capacity ($\frac{\text{Gbit}}{\text{s}}$)	ISL_{cap}
Satellite Lifetime (years)	l	Solar Efficiency Decay Rate	$r = 0.03$ [41]
Heat Pump Coefficient of Performance	$CoP = \frac{T_{amb}}{T_{rad} - T_{amb}}$ [12]	Heat Pump Power (W)	$P_{hp} = \frac{P_{comp}}{CoP}$
ISL Power (W)	$\frac{ISL_{cap}}{3}$ [58]	EOL Power (W)	$P_{EOL} = P_{comp} + P_{hp} + P_{ISL} + P_{rest}$
BOL Power (W)	$P_{BOL} = P_{EOL} \times (1-r)^{l}$	Power Subsystem Mass (kg)	$m_P = 4.384 \times 10^{-3} P_{BOL} - 3.5 \times 10^{-1}$
Structure Mass (kg)	$m_{struc} = 9.384 \times 10^{-3} P_{BOL} + 221$	ADCS Mass (kg)	$m_{adcs} = \frac{P_{BOL}}{26.65}$
C&DH Mass (kg)	$m_{c\&dh} = 2.52 \times 10^{-2} P_{BOL} + 1.25 ISL_{cap}$	Fuel Mass (kg)	$m_{fuel} = \left[e^{\frac{\Delta v}{v_{exhaust}}} - 1\right] m_{dry}$ [56]

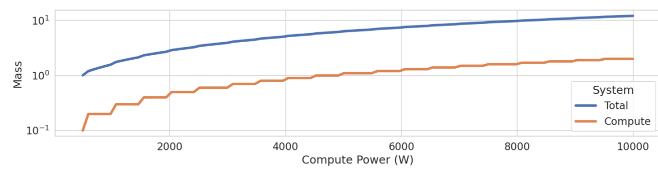


Figure 6: Compute power vs mass. Masses are relative to the total mass of a 500 W SµDC.

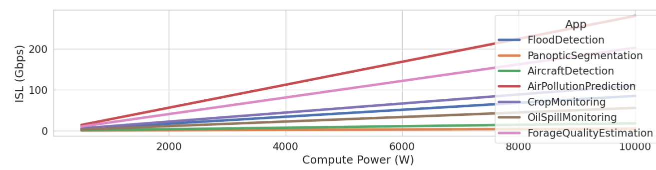


Figure 8: ISL datarates required to saturate RTX 3090 GPUs for several satellite imagery applications taken from [9], with compute power in 0.5 kW to 10 kW.

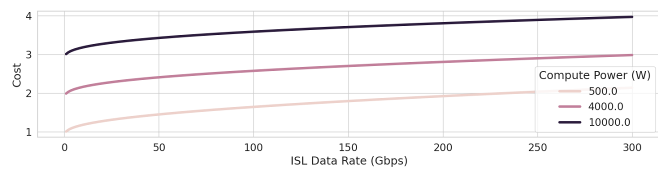


Figure 7: TCO vs Communication.

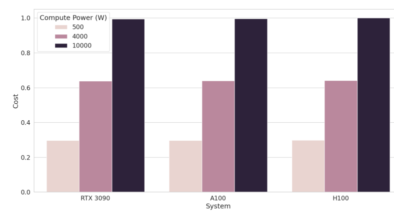


Figure 9: TCO vs Architecture

conclusions apply to other processor architectures as well). Based on these results, we see that a 500 W SµDC needs no more than 25 $\frac{\text{Gbit}}{\text{s}}$ ISL to support even the most lightweight applications, which corresponds to a less than 30% increase in TCO. Ensuring sufficient ISL capacity for 4 kW and 10 kW SµDCs is even more affordable — both see less than 26% TCO increase to support ISLs sufficient for the most lightweight applications. In reality, ISL requirements and, therefore, impact on TCO will be much lower as some applications require significantly more computation, and thus less channel capacity is needed to saturate compute. Ongoing improvements in FSO power efficiency promise to further decrease the TCO impact of ISLs [42], [70]. Overall, the results show that **the impact of communication on SµDC TCO is small**.

[2]Estimated based on # of DSPs

TABLE II: Price, TDP, and TFLOPs for several GPGPU architectures, and several radiation hardened architectures. Data on radiation hardened designs is from [72].

System	TID (krad(Si))	Price ($)	TDP (W)	TFLOPs (FP32)	TFLOPs (TF32)
RTX 3090	2 to 10	1690	350	35.58	N/A
A100	2 to 10	17 210	300	19.5	156
H100	2 to 10	43 989	350 [67]	51	756
Radeon 780M	2 to 10	N/A	15	8.29	N/A
BAE RAD750	200	200 000	5	0.000 27	N/A
MPC8548E	100	200 000	5	0.008	N/A
Virtex-5QV	1000	75 000	15	0.08	N/A
Kintex UltraScale XQR	100	N/A	N/A	0.65[2]	N/A

Then we focus on the impact of processing hardware architecture (Table II) for a given compute power budget. Note that the A100 and H100 have max $\frac{\text{FLOPs}}{\text{W}}$ advantage of 5.1× and 21.2×, respectively, over RTX 3090. The A100 and H100 achieve this high power efficiency via inclusion of 'TensorCore's which accelerate the low precision (TF32) tensor-based arithmetic found in DNNs. But their max $\frac{\text{FLOPs}}{\$}$ are worse - 0.50× and 0.82× than the RTX 3090. Thus, a terrestrial datacenter may choose to use a large number of RTX 3090 systems rather A100 or H100 since better $\frac{\text{FLOPs}}{\$}$ of RTX 3090 will reflect as better $\frac{\text{FLOPs}}{\$_{\text{TCO}}}$ - in terrestrial datacenters, server costs up to 72% of TCO, while power cost is < 10% [8], [30]. For a SµDC, however, compute costs are only a tiny fraction of the overall TCO, and A100 (H100) achieve > 6× (> 9×) better energy efficiency on EO applications than RTX 3090 [9].

Figure 9 shows TCO dependence across architectures. We see that the TCO effects are minimal due to relatively low cost of the compute. This means that A100 and H100 are a lot more attractive than RTX 3090 since, unlike the case of terrestrial datacenters, they will provide much higher $\frac{\text{FLOPs}}{\$_{\text{TCO}}}$ for SµDCs (as it is power that greatly affects the overall TCO). In general, **architectures with the highest $\frac{\text{FLOPs}}{\text{W}}$ provide much higher performance per TCO \$ for a SµDC, even if they have poor $\frac{\text{FLOPs}}{\$}$.**

Finally we study the impact of compression on TCO since compression can be used to decrease communication costs. Figure 10 gives TCO cost scaling of a 4 kW SµDC using different compression algorithms. As this does not include power cost of decompression, these are upper bounds on the possible TCO improvements. With RTX-3090 servers, CCSDS[3] provides $< 3\%$ TCO savings, lossless JPEG2000 provides 5% TCO savings, and a high PSNR, quasi-lossless neural compression algorithm [7] provides 8% TCO savings. However, as compute hardware becomes more energy efficient, the portion of overall cost determined by the ISL grows. Thus, asymptotically, **the compression algorithms provide** 11.7%, 20.5%, and 26.5% **decreases in TCO as the energy efficiency scaling grows**.

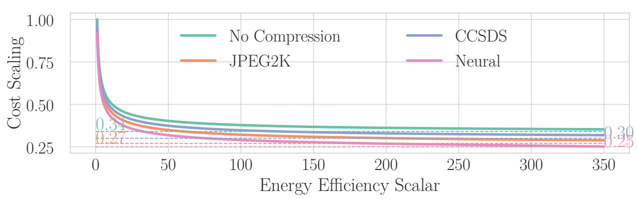


Figure 10: TCO vs Energy Efficiency for a 4 kW SµDC using different compression algorithms to reduce ISL capacity requirements.

A. *Power impact on SµDC vs terrestrial datacenter TCOs*

Fig. 11 shows normalized TCO for two satellite TCO models and three terrestrial datacenter TCO models without NRE amortization. In the satellite context, "[Infra]structure" refers to the satellite bus structure, while in the terrestrial context, it refers to datacenter facilities. Similarly, "Networking" refers to off-satellite communications (ISL, downlinks) in the satellite context, while it refers to inter- and intra-datacenter networking costs in terrestrial datacenters. We see that TCO for terrestrial datacenters is dominated by server and facilities costs, not power, while power dominates TCO for SµDCs.

[3]a standard lossless compression algorithm for use in space

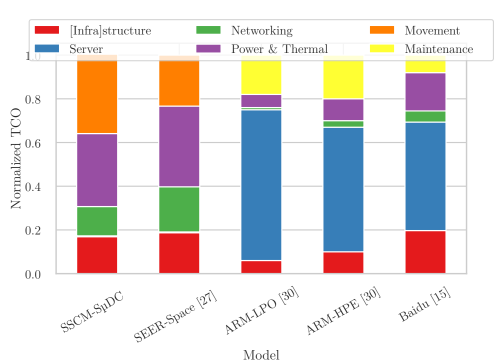


Figure 11: Normalized datacenter TCO from two satellite TCO models and three terrestrial datacenter TCO models.

There are two key reasons why power constitutes a significant fraction of TCO for space datacenters whereas it is only a small portion in the terrestrial datacenters. First, the cost of power is much lower for terrestrial datacenters. Terrestrial datacenters typically draw power from a grid whose capital and infrastructure costs get amortized over a large number (often millions) of commercial and non-commercial users, receive significant government tax credits and other subsidies that reduce their cost of power (even the grid is typically heavily subsidized), and pay only for the actual power used (instead of the worst case). Space datacenters, on the other hand, generate their own power and directly pay for it (instead of drawing power from a grid), do not share their power source and, therefore, have no amortization benefits, do not receive any governmental support to reduce power cost, and have to allocate power generation for the worst-case (since there is no grid to draw excess power from). Second, power in a space datacenter impacts the cost of other components in ways that are unique to a space datacenter, magnifying the impact of power on the total cost. For example, the mass of the satellite grows directly with the size of the solar panel and the size of the radiator. This means that thruster rocket fuel must increase proportionally with the increase in satellite dry mass. In addition, larger and more expensive thrusters may be needed to produce proportionally more thrust. All of this increases overall cost

B. *Thermal Management*

Since satellites are in vacuum, the only way heat escapes from a satellite is via radiation — convection and conduction may move heat within the satellite but cannot move heat away from the satellite. Equation 1 shows the heat radiated by the satellite as a function of surface area incident to free space (A), and the emissivity (ϵ) and the temperature (T) of the satellite (vs the temperature of the background).

$$P_{\text{emit}} = \epsilon \sigma A_{\text{space}} T^4, \tag{1}$$

where The Stefan-Boltzmann constant is $\sigma = 5.67 \times 10^{-8} \frac{\mathrm{W}}{\mathrm{m^2\,K^4}}$.

Radiation is extremely efficient in space due to the 2.7K temperature of the space background (vs 270K+ on earth). A $1\,\mathrm{m}^2$ radiator ($\epsilon = 0.86$) at $45\,^\circ\mathrm{C}$ will emit just shy of 1 kW when both radiator faces are oriented toward deep space. Only a $4m^2$ radiator can support the heat dissipated by our 4 kW SµDCs. When required, the radiator temperature is increased using an active thermal control system [6], [54], [81] to increase the amount of heat dissipation even further. Radiators on satellites with active cooling are routinely used to remove heat in excess of 10 kW [6], [54], [81]. Figure 12 shows the trade-offs between radiator size and radiator temperature needed to achieve fixed amounts of emitted power (radiative flux).

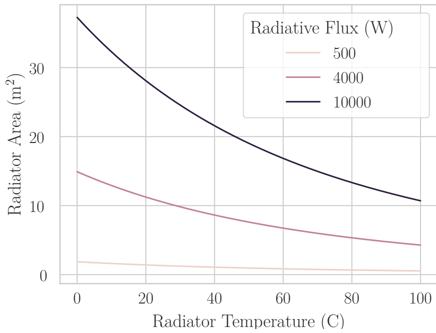


Figure 12: Radiator size vs temperature. The curves show the required panel radiator area needed to radiate 500 W, 4 kW, and 10 kW when both panel faces point towards empty space. Radiators have emissivity $\epsilon = 0.86$ [92].

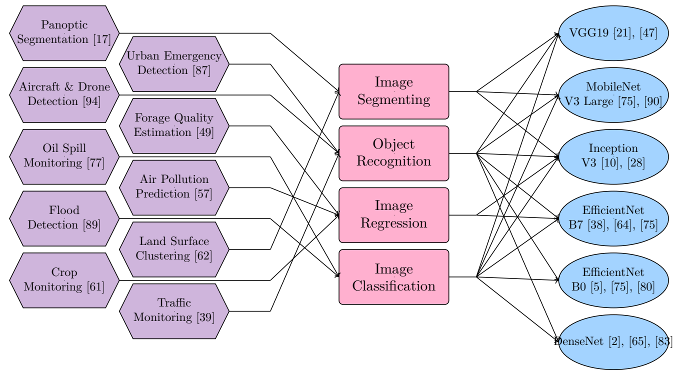


Figure 13: Applications, tasks, and kernels.

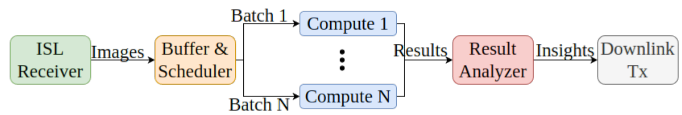


Figure 14: The payload performs batch-computing of EO satellite imagery.

IV. Extreme Heterogeneity for Low-cost In Space Computing

A. Applications

With the above insights, we ask the question - what are good potential architectures for a SμDC's compute payload to minimize TCO? To address this question, we consider streaming, non-longitudinal applications rather than longitudinal applications which require storing large datasets. Bleier *et al.* [9] identified a sampling of non-longitudinal applications processing EO satellite imagery. Users of these applications include NASA [82], ESA [68], the California Air Resources Board [11], USDA [26], the Ministry of Agriculture of China [97], the US Department of Transportation [51], etc. These applications perform object recognition, image classification, image regression, and image segmentation tasks on the satellite data. For these tasks, artificial neural networks — predominantly convolutional neural networks (CNNs) — have emerged as the computational kernel of choice, due to their high accuracy and precision. Figure 13 depicts the relationship between these applications, the image processing tasks, and the CNNs which have been deployed for these tasks in the context of satellite data image processing.

We first consider GPU-based data processing. Our EO image processing workloads fit fully into a single GPU. As such, expensive, and high power interconnects, such as those found in NVIDIA DGX servers, are not necessary. Instead, discrete GPUs can compute batches in parallel, as depicted in Figure 14. Batching may induce some latency between image generation and image processing. A LEO Earth observation satellite may produce around six images per minute (exact rate depends on orbital velocity, and ground frame size), and a SμDC may receive images from one or more observation satellite. Thus, it may take up to several minutes for an energy-minimizing batch size to be reached. In this scenario, a suboptimal batch size may be used. In addition, this latency is still significantly better than the latency achieved using a traditional bent-pipe downlink model [19]. As compute completes, the results are sent to an analyzer, which determines whether the results are 'insights' which should be downlinked to Earth, or whether the results contain little relevant information, in which case they can be discarded. We focus on a 4 kW SμDC. This size allows a single SμDC to support constellations of 64 EO satellites at current imaging resolutions for nearly all applications, as shown in Tab. III.

Table III depicts performance and energy characteristics of these application workloads on an RTX 3090 GPU — a commodity class GPU manufactured in Samsung's 8 nm tech node. For these results, we consider offline batch processing of workloads since many EO image processing applications are latency insensitive — current EO image processing latencies are measured in hours [86], due in large part to the time it takes an LEO satellite to orbit above a downlink station. Also, batch processing is more energy efficient than latency sensitive online or stream processing, since it enables utilizing energy-minimizing batch sizes, and the lack of work-item level latency constraints means high power operating points can often be avoided [98].

As we see, the commodity GPU shows poor energy efficiency, which, as shown earlier, leads to high TCO costs (since higher compute power is needed for processing). Below we consider accelerator-based architectures to improve energy efficiency and lowering space microdatacenter TCO.

B. Design Space Exploration

Our TCO analysis (Section III) showed that architectural optimizations focused on energy efficiency can have

TABLE III: Application performance on RTX 3090 commodity GPU. The number of 4 kW SμDCs with RTX 3090 GPUs needed to support a constellation of 64 EO satellites is shown in the rightmost column.

App Name	P (W)	Util (%)	Infer time (s)	$\frac{\text{kpixel}}{\text{J}}$	# SμDC
Air Pollution	119	25	0.59	1168	1
Crop Monitoring	222	42	1.57	395	1
Flood Detection	325	88	5.53	307	1
Aircraft Detection	124	26	0.26	74	1
Forage Quality Estimation	129	27	0.56	843	1
Urban Emergency Detection	266	72	2.04	569	1
Oil Spill Monitoring	347	98	3.84	231	1
Traffic Monitoring	19	< 1	2.72	2597	1
Land Surface Clustering	108	2	0.35	2175	1
Panoptic Segmentation	160	80	7.81	20	4

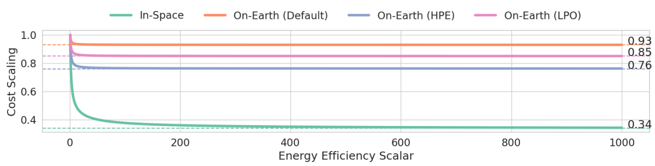


Figure 15: Relative TCO vs Energy Efficiency for in-space and terrestrial data centers, assuming compute hardware costs are invariant.

substantial impact on TCO for SμDCs. Figure 15 shows TCO scaling for in-space and terrestrial datacenters as hardware energy efficiency improves, assuming 1) baseline of commodity hardware, and 2) hardware costs remain constant. TCO breakdown for terrestrial datacenters is from [30]. 'On-Earth (Default)' assumes that only energy costs scale. The 'On-Earth (HPE)' and 'On-Earth (LPO)' curves also scale the cost of in-datacenter power distribution for high performance and low-power, high-density server configurations, respectively. According to this model, server costs range from 57% to 72% of TCO, while power costs are only 7% to 13% of TCO in terrestrial datacenters. This is very different from TCO breakdown of a SμDC where $< 1\%$ of TCO is in computer hardware, and over a third of TCO is in power and thermal management subsystems. This has two implications — 1) changes to hardware cost have large impact on terrestrial TCO, but only minor impact on SμDC TCO, and 2) improvements in energy efficiency may have large impact on SμDC TCO, but muted impact on terrestrial datacenter TCO unless coupled with decreased hardware cost[4].

We see that the impact of compute energy efficiency on TCO of a terrestrial datacenter is minimal - less than ten percent for the On-Earth (Default) case. Even when accounting for on-premise power distribution hardware, the impact of compute energy efficiency on TCO is limited to twenty-five percent (On-Earth (LPO)). In space, however, increased energy efficiency of compute leads to a nearly sixty-six percent decrease in TCO. Thus, when not accounting for compute hardware cost, TCO for an in-space datacenter is up to 250% more sensitive to energy efficiency improvements than terrestrial datacenters.

[4]Low-cost energy efficient hardware such as neural network accelerators may still be passed over for relatively expensive and power consuming, general purpose hardware if software cannot easily be ported to the accelerators.

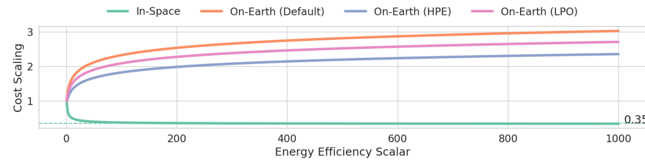


Figure 16: Relative TCO vs Energy Efficiency for in-space and terrestrial data centers, assuming compute hardware costs scale logarithmically with the energy efficiency scalar.

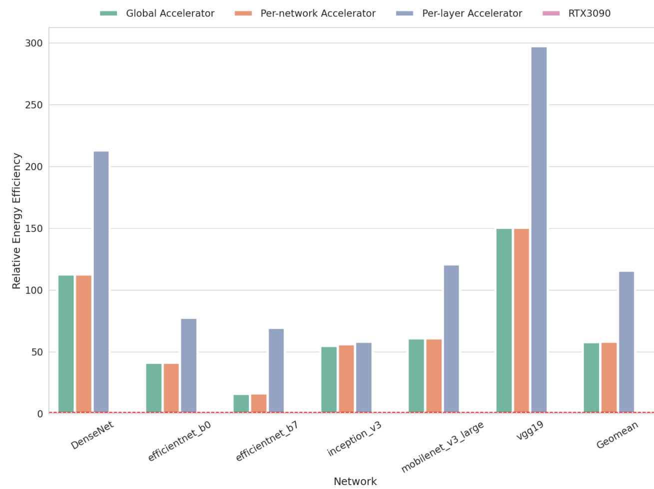
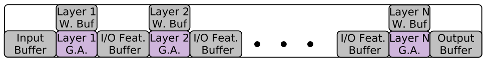


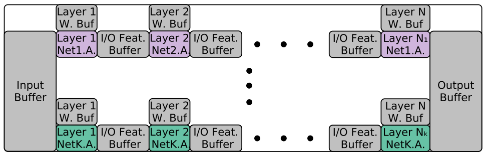
Figure 17: Energy efficiency improvements for accelerator architectures vs baseline commodity GPU.

In reality, energy efficiency improvements rarely come for free — they often impact cost. Figure 16 assumes logarithmic price scaling of compute hardware with respect to energy efficiency. Thus, for example, computer hardware which is 100× more energy efficient than baseline costs 3× more money. Even with this highly sublinear price scaling, TCO for terrestrial datacenters increases dramatically — over a 100% increase in TCO with 200× energy efficiency scaling. This is because compute hardware, not energy, make up the majority of terrestrial datacenter TCO. In space, however, the price of commodity hardware has minimal impact on TCO. Thus, this sublinear price scaling is effective in space. Even at a 1000× energy efficiency scalar, TCO is still decreasing.

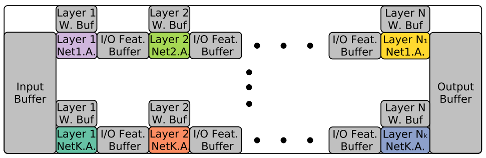
This suggests that architectures that trade off cost for energy efficiency may be much better fit for space. Figure 17 demonstrates the achievable energy efficiency gains from a limit study using accelerator-based architectures over commodity GPUs on the neural networks listed in Figure 13. The three accelerator system architectures considered are: 1) *Global Accelerator* — a homogeneous system using the accelerator with the best geometric mean energy efficiency across all networks, 2) *Per-Network Accelerator*



(a) A homogeneous global accelerator architecture.



(b) A heterogeneous per-network accelerator architecture.



(c) A heterogeneous per-layer accelerator architecture.

Figure 18: Three accelerator-based design points replace the GPU compute blocks of Fig. 14 with one or more different type of accelerator. In each design, distinct accelerator designs have distinct colors.

– a heterogeneous system of accelerators consisting of the best accelerators for each network, 3) *Per-Layer Accelerator* – a heterogeneous system of accelerators consisting of the best accelerators for each layer. These designs are depicted in Figure 18. They replace the 'Compute k' blocks of Figure 14 with a pipeline of ANN accelerators. In Figure 18a, a single accelerator design is used to run all workloads. In Figure 18b, each individual network has its own accelerator. In Figure 18c, each layer of each network has its own accelerator. For each accelerator pipeline, each layer's output features are double buffered in the I/O feature buffers, enabling asynchronous pipelined execution. In the heterogeneous architectures, inputs to a neural network are dispatched from the Input buffer to one of the several parallel pipelines which corresponds with the selected neural network. Neural network outputs are buffered in the output buffer before being sent to the Results Analyzer for insight extraction.

In each design, outputs of each non-final layer are double buffered as inputs to the subsequent layer, enabling pipelined execution. Since, in all cases, only a single neural network is executed at a time, the per-network and per-layer accelerator designs have a unified input buffer shared by each layer-1 accelerator, and an output buffer shared by each final layer. Weights and biases for each layer are stored in layer-specific weight buffers.

Energy values for the accelerator designs are estimated using the Timeloop-Accelergy framework [95]. Timeloop-Accelergy was also used to perform a design space exploration over Eyeriss-like [13] accelerators using row-stationary dataflows to identify the best Global, Per-Network, and Per-Layer accelerators. Dimensions in the design space exploration are the length of the PE grid in x and y dimensions and the size of input feature, weight, and accumulation buffers. A total of 7168 designs were evaluated. In order to determine the globally optimal (energy minimizing) design, we use a geometric mean of each design's energy efficiency on all neural network layers. Similarly, to determine the per-network optimal design, we use geometric mean of each design's energy efficiency on all layers of the network. Energy values for the GPU baseline were evaluated on a RTX-3090. Evaluations used CUDA version 11.7, cuDNN version 8.9.0, and TensorFlow version 2.12. To find the most energy efficient batch sizes, we ran inference 100 times on different batch sizes, and used Python NVML to measure the average GPU utilization and power consumption.

We see that the Global Accelerator system provides an average 57.8× improvement to energy efficiency over the baseline — sufficient to achieve a 60% reduction in TCO. Heterogeneous architectures provide up to 116× on average — sufficient to achieve a 63% reduction in TCO.

In general, the following insights emerge about architectures for space microdatacenters. First, energy-efficient hardware holds promise to provide TCO savings higher than would be expected from energy efficiency alone, because mass reductions in power supply and thermal management subsystems also lead to lower cost for other satellite subsystems (e.g., propulsion, ADCS, structure, etc.). Second, accelerator-based architectures are highly effective in reducing the TCO of SµDC even with a higher monetary cost for the accelerators. This is because of the high cost of energy systems (i.e., power generation and thermal management) relative to the compute hardware for spacecraft. This stands in stark contrast with terrestrial datacenters, where even logarithmic cost scaling of energy efficient hardware leads to doubling of TCO, since the majority of TCO for such datacenters is in hardware costs, and only a small portion of TCO is in energy. Third, homogeneous accelerator systems provide sufficient energy efficiency to capture nearly all possible TCO reductions for in-space datacenters, given current ISL power consumption. However, as ISLs continue to improve in bandwidth and energy efficiency (e.g., via DARPA's Space-BACN project [16]), the TCO benefits of heterogeneous accelerator architectures over homogeneous accelerator systems may increase (since the TCO associated with ISLs may drop, which in turn increases the relative cost of compute). Compression may also be used to reduce the impact of ISL cost.

V. Collaborative Compute Constellation Architecture

The constellation architecture that all our prior analysis assumed is one where EO satellites lack significant computational capabilities. I.e., we assumed (Fig. 20a) that the EO satellites are unable to run applications, or perform data filtering — they only offload data to a SµDC for processing.

However, there is significant interest in satellite edge computing. For example, prior works have considered using compute on EO satellites to perform data cleaning — filtering out unusable images, such as those occluded by clouds [18], [19] on CubeSat class EO satellites. By performing this filtering, the amount of data needed to be downlinked to Earth is reduced. While these works have attempted to directly address the 'downlink deficit', this filtering may instead be applied to reduce the amount of data sent via ISL to SµDCs. This allows reducing the total amount of data which must be transmitted to a SµDC, and thus also the total amount of processing a SµDC must perform. This enables savings by reducing the SµDC's ISL capacity and power requirements, as well as power dedicated to computation and thermal management. Such a constellation architecture - collaborative compute constellation - is depicted in Figure 20b, in which the EO satellites leverage their own compute hardware to reduce data transmission to the SµDC.

Figure 19 shows the TCO of a SµDC required to support a constellation as edge filtering rates improve. The decrease in cost is due to the shrinking size of the required SµDC. At a filtering rate of zero, a 4 kW SµDC is required, but at a filtering rate of 0.5, only a 2 kW SµDC is required.

Figure 21 shows the sensitivity of TCO benefits to the energy efficiency factor of the compute hardware (normalized against the TCO of a 4 kW SµDC). Since a collaborative constellation reduces SµDC ISL and compute power proportionally, it may be especially attractive for SµDCs equipped with energy efficient, heterogeneous architectures. Assuming cloud filtering (resulting in $\approx \frac{2}{3}$ reduction in data transmitted), a collaborative compute constellation architecture provides a $1.74\times, 1.33\times$, and $1.31\times$ improvement in TCO against baseline SµDCs with a commodity GPU-based architecture, a global accelerator architecture, and a heterogeneous architecture, respectively, for a 4 kW SµDC baseline.

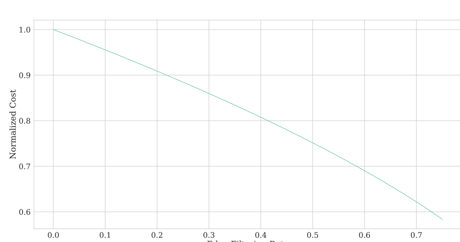


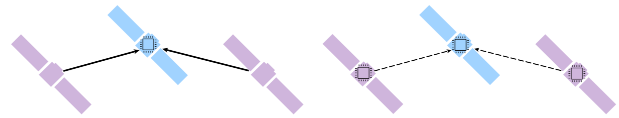
Figure 19: Relative TCO vs edge satellite filtering rate. Baseline is a 4 kW SµDC.

VI. Distributed vs. Monolithic Architecture

A. *Wright's Law and TCO*

Our earlier TCO analysis did not consider experience effects. In aerospace, as well as many other manufacturing sectors, the impact of learning on unit costs is modeled via Wright's Law [60]: $C_n = C_1 \cdot n^{\log_2(b)}$, where C_n is the cost of the nth unit, and b is the 'progress ratio'. Wright's law states that every time the number of units manufactured doubles, the cost of the next unit will have gone down by a fixed percentage. For example, if $C_1 = \$1$, and $b = 0.9$, then $C_2 = \$0.90$, and $C_4 = \$0.81$, etc.

Wright's law is especially powerful in aerospace, where it originated [93]. Since spacecraft are very complex to manufacture, a high progress ratio can be achieved — often in the $b \in [0.7, 0.8]$ range [32], [52]. The experience effect is most profound in component manufacturing and in satellite assembly [31]. Figure 22 shows the impact of Wright's law on marginal cost for several SµDC design points assuming $b = 0.75$. The initial satellite design, which includes NRE costs, is high, but the marginal costs quickly decrease. By the time the 100th satellite is manufactured, cost has decreased by over 50%. In fact, the 100th 10 kW SµDC is cheaper than the first 4 kW SµDC. Wright's Law has also been applied *to* NRE. A recent review of F-15 fighter jet procurement shows that experience models such as Wright's Law predicted decreases in research and development costs for advanced models of the F-15 [55].



(a) A baseline constellation configuration. (b) A collaborative compute constellation configuration.

Figure 20: In the baseline configuration (a), all data generated by EO satellites must be transmitted to the SµDC. By performing filtering on the edge (b), EO satellites reduce the amount of data they have to send to the SµDC, which reduces the communication and compute power consumption of the SµDC.

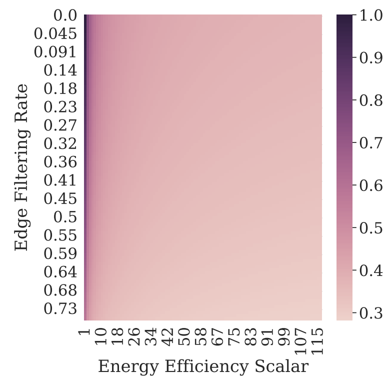


Figure 21: Normalized TCO vs energy efficiency of compute hardware and filtering capability of the EO satellites.

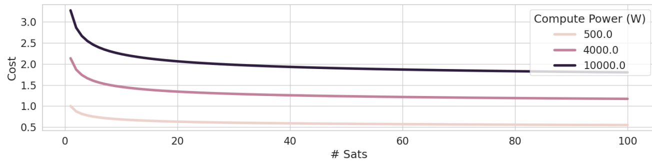


Figure 22: Satellite marginal cost vs number of satellites.

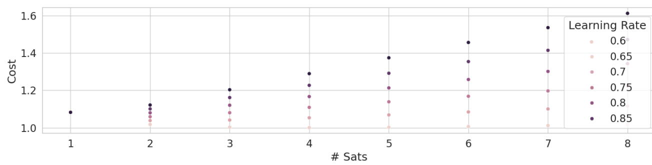


Figure 23: TCO (NRE and RE) vs # of satellites in constellation with fixed target of 32 kW.

B. A TCO Case for Distributed Space Microdatacenters

In this work we motivate distributed in-space computing using multiple SµDCs by quantifying its effect on TCO. We will show that by using multiple small SµDCs, total cost can be reduced relative to a monolithic, large SµDC.

The above results suggest that the total cost of ownership for a constellation of SµDCs increases sublinearly with the size of the constellation. It is then worth asking: what is impact on TCO of using multiple small SµDCs, relative to a monolithic, large SµDC. I.e., to reach a target compute power (32 kW, for example), should we build a single 32 kW SµDC, 2× 16 kW SµDC, 3× 10.66 kW SµDC, etc, to minimize the TCO?

Figure 23 shows the results for different values of the Wright's law progress ratio, or learning rate. NRE is included in this analysis, and is thus amortized across the number of satellites manufactured. For a pessimistic progress ratio (0.85), a monolithic system minimizes TCO. However, for all other progress ratios, a distributed system of multiple SµDCs minimizes TCO. With an optimistic ratio ($\leq$ 0.65 — meaning costs scales by 0.65 for every doubling of production), TCO is minimized at greater than 4 SµDCs, and with TCO over 10% below a monolithic system.

VII. Near-Zero Cost Overprovisioning

Our TCO analysis (Figure 5) showed that the monetary cost of compute as a fraction of TCO was insignificant ($<$ 1%). This suggests that compute can be overprovisioned in a SµDC at near-zero cost (as long as the excess compute is kept powered off). This overprovisioning can then be used to enhance a SµDCs' availability.

Figure 24 shows that overprovisioning increases the likelihood of full system availability. We model the lifetime of each physical compute node as $X_i \sim \text{Exp}(\lambda)$ where $T = \frac{1}{\lambda}$ is the mean time to failure. Assuming homogeneous compute nodes, the X_i are independent and identically

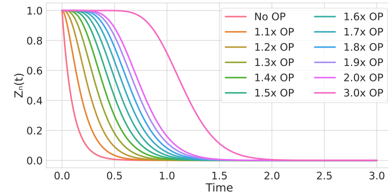


Figure 24: The likelihood that at least 10 servers work vs time for overprovisioning factors.

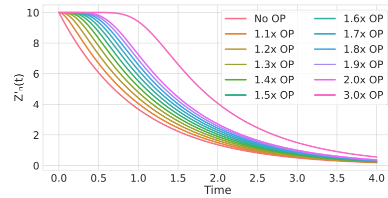


Figure 25: The expected number of working servers vs time (capped at 10 due to power limits).

distributed. Let $Y_i(t)$ be the indicator function such that $Y_i(t) = 1$ if $t < X_i$, else 0.

Let Z_n be a parametric family of continuous time random processes for $n \in \mathbb{N}$ with $n \geq 10$ such that $Z_n(t) = 1$ if $\sum_{i=1}^{n} Y_i(t) \geq 10$, else 0.

Figure 24, then, shows the probability that $Z_n(t) = 1$, i.e., that at least ten physical nodes are still working, for choices of $10 \leq n \leq 30$. We note two main takeaways: first, the median time to system degradation, that is < 10 nodes working, increases *superlinearly* with overprovisioning factor for small values of overprovisioning factor. With ten physical nodes, the median time to system degradation is $0.25T$, but with 20 and 30 physical nodes, the median time to system degradation is $0.8T$ and $1.25T$ respectively. Second, the time at which probability of system degradation exceeds 99% also grows superlinearly: 0.46, 1.43, and 1.89 for $n = 10, 20,$ and 30, respectively. Thus, at near-zero cost, compute overprovisioning substantially increases the time SµDCs will operate at full capacity.

Overprovisioning also assists in graceful degradation of a SµDC's capabilities. Let Z'_n be parametric family of continuous time random processes for $n \in \mathbb{N}$ with $n \geq 10$ such that $Z'_n(t) = \min\{10, \sum_{i=1}^{n} Y_i(t)\}$. That is, $Z'_n(t)$ represents the number of compute nodes which are usable by the SµDC at time t. Figure 25 depicts $\mathbb{E}[Z'_n(t)]$ for choices of $10 \leq n \leq 30$. This shows that at all times, overprovisioning provides significant improvement in the expected computational power of a SµDC, even if this amount is less than the BOL maximum.

In addition, near-zero cost overprovisioning can be used for lifetime management. The computer hardware in a SµDC can include both accelerators which are guaranteed to be useful at the beginning of the satellite's lifetime, and general purpose, programmable computing systems. As the satellite ages, and new applications emerge which cannot be effectively supported on the accelerators, computing is provided by the general purpose architectures. This approach is enabled by the unique cost breakdown of a SµDC, in which hardware capital costs are low relative to overall TCO.

VIII. Reliability Implications

All our previous analysis assumed COTS hardware, not radiation-hardened hardware. COTS hardware is strongly preferable for SµDCs since radiation-hardened hardware can have prohibitive costs. The bottom four rows of Table II list four radiation hardened processors. As an illustrative example of the high costs of rad-hard hardware, the rad-hard Virtex-5QV FPGA is 27× less energy-efficient than H100 in an IEEE FP32 comparison. It is 405× less efficient if the H100 utilizes its tensor cores with TF32 support.

The question then becomes: can COTS hardware meet the reliability requirements of SµDCs? Or must we use rad-hard hardware?

Radiation related effects on hardware are categorized into long term, exposure based effects, also called total ionizing dose (TID), and single event effects. As shown in Table II, radiation hardened designs provide protection for TIDs $\geq 100\,\text{krad(Si)}$. This is especially useful for satellites in GEO — such satellites often have long mission durations and are located inside the outer van Allen radiation belt. Computers in GEO protected by 200 mils of aluminum shielding expect to see $4\,\frac{\text{krad(Si)}}{\text{yr}}$ [71]. However, satellites in non-polar LEO orbits typically see only $\sim 0.5\,\frac{\text{krad(Si)}}{\text{yr}}$ with 200 mils of aluminum shielding, and this can be reduced to only $\sim 0.2\,\frac{\text{krad(Si)}}{\text{yr}}$ with 400 mils of shielding [48]. Further, LEO satellite lifetimes are often short — due to atmospheric drag, LEO orbits will decay if satellites do not periodically perform rocket burns. As such, lifetimes are limited by the amount of stored rocket fuel. For example, the LEO satellites of the Starlink constellation (which make up a majority of all artificial satellites in orbit), target an operational lifetime of five years [78].

Fortuitously, TID tolerance of mainstream (i.e., non-rad hard) commercial technologies has been increasing with technology scaling. Figure 26 shows TID tolerance for several processors at different tech nodes. At 14 nm tech node, processors can tolerate an order of magnitude more radiation than would be experienced during an LEO satellite's lifetime. Newer commercial technologies provide even higher resilience against permanent, destructive failures caused by high energy particles [23].

As such, LEO satellites are increasingly turning to non-radiation hardened computers due to the high cost, poor performance, relative scarcity of radiation hardened components, and the more permissive radiation environment of LEO. For example, Starlink satellites use a COTS Xilinx SoC for their GDGPS navigation subsystem [14], Dragon and Falcon 9 spacecraft use non-radiation hardened x86 dual-core machines for their flight computers [20]

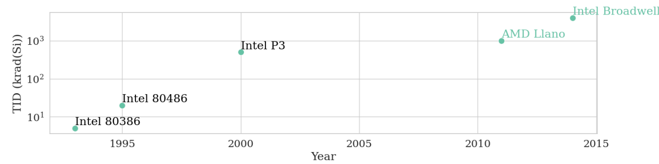


Figure 26: Total ionizing dose before failure [34], [36], [44], [74], [79] (no failures for Intel Broadwell and AMD Llano).

(and in fact, use no radiation hardened computers at all, despite being safety critical systems), the ESA's ϕ-Sat-1 hyperspectral imaging satellite uses an Intel Movidius Myriad 2 VPU [29]. Further, many mission critical satellite instruments use COTS hardware [24], [25]. Among many other COTS computers, the ISS hosts the HPE Spaceborne Computer 2 [35]. On Mars, the Perseverance rover uses redundant Intel Atom SBCs (2× COMEX-IE38s) for its image processing and data compression requirements of its 23× cameras, while it uses a BAE RAD750 for critical tasks [84]. A similar design methodology is used in commercial aircraft — highly reliable, low performance flight control systems are paired with relatively high-performance multimedia processors for passenger entertainment systems. This split design also makes sense for SµDCs — radiation hardened systems for flight control, and non-radiation hardened systems for its application processing payload. This is the reason COTS hardware was assumed throughout our analysis.

Fig. 28 shows the impact of different reliability schemes (triple modular redundancy (TMR), dual modular redundancy (DMR), and software-based redundancy) on TCO for equivalent computing power between 0.5 kW to 4 kW. For TMR and DMR, we assume an overhead of 3× and 2× respectively. A DMR scheme at 2 kW equivalent computing power, for example, is assumed to consume $\sim 4\,\text{kW}$. For software, we assume an overhead of 20%. This is conservative - ANNs are remarkably resilient to soft errors. Fig. 27 shows the impact of soft errors on image classification using several different ANNs on ImageNet. This makes a number of pessimistic assumptions, including that all soft-errors result in an incorrect inference, and that soft-errors never result in a correct inference. Previous work on hardening ANNs [3], [73], [76], [88] has reported no more than 20% overhead. The results show that impact of hardware redundancy-based solutions on SµDC TCO can be high (again due to the impact also on power generation and thermal subsystems). Software-based reliability solutions have small cost in terms of TCO.

IX. Summary and Conclusions

This paper presented a system-level approach to architecture of server-based computing systems in space (SµDCs). We extended the SSCM cost model to include computer hardware and support system costs for SµDCs, revealing compute power as the primary TCO determinant

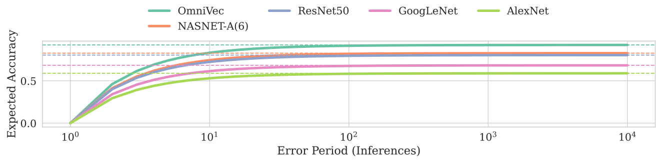


Figure 27: The impact of soft-errors on ImageNet.

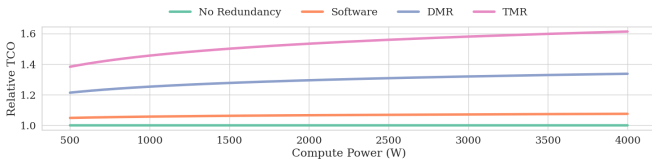


Figure 28: Relative TCO for different redundancy choices for various levels of equivalent computing power.

with sublinear dependence. Factors like compute mass & monetary cost, and communication have minor impact on TCO. We showed that special-purpose, accelerator-based architectures offer significant TCO advantages in SµDCs compared to general-purpose architectures. Specifically, the use of extremely heterogeneous designs (e.g., one accelerator per layer of a neural network) can reduce SµDC TCO by 116× in spite of their poor $\frac{\text{FLOPs}}{\$}$ characteristics. In addition, we showed that the use of collaborative compute constellations — constellations in which EO satellites are also capable of doing some rudimentary data filtering — further improves SµDC TCO by 1.31 to 1.74×, a distributed architecture reduces TCO by 10% over a monolithic architecture, and low monetary cost of compute can be leveraged to provide near zero cost compute overprovisioning which improve an SµDC's availability significantly and support graceful degradation. Overall, this is the first paper on cost-aware architecture and optimization of a SµDC.

Acknowledements

The authors would like to thank Sam Sanchez and Chris Hutchings from Gallorath, Eric Mahr and Alex Duvall from Aerospace Corporation, and Dave Bearden from JPL for their consultation regarding the SEER-Space and SSCM cost models. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions which have improved the quality of this paper.

References

[1] M. Aazam and E.-N. Huh, "Dynamic resource provisioning through fog micro datacenter," in *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 2015, pp. 105–110.

[2] S. R. Abdani and M. A. Zulkifley, "Densenet with spatial pyramid pooling for industrial oil palm plantation detection," in *2019 International Conference on Mechatronics, Robotics and Systems Engineering (MoRSE)*, 2019, pp. 134–138.

[3] G. Abich, J. Gava, R. Garibotti, R. Reis, and L. Ost, "Applying lightweight soft error mitigation techniques to embedded mixed precision deep neural networks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 11, pp. 4772–4782, 2021.

[4] Aerospace Corp. (2019) Small satellite cost model. [Online]. Available: https://aerospace.org/sscm

[5] T. Ahmed and N. H. N. Sabab, "Classification and understanding of cloud structures via satellite images with efficientunet," *SN Computer Science*, vol. 3, no. 1, p. 99, 2022.

[6] (2008) Alos user handbook. [Online]. Available: \url{https://www.eorc.jaxa.jp/ALOS/en/doc/fdata/ALOS_HB_RevC_EN.pdf#page=45}

[7] P. Bacchus, R. Fraisse, A. Roumy, and C. Guillemot, "Quasi lossless satellite image compression," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 1532–1535.

[8] L. Barroso, U. Hölzle, and P. Ranganathan, *The datacenter as a computer: Designing warehouse-scale machines.* Morgan and Claypool, 2009.

[9] 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, p. 900–915. [Online]. Available: https://doi.org/10.1145/3613424.3614271

[10] K. Boonyuen, P. Kaewprapha, U. Weesakul, and P. Srivihok, "Convolutional neural network inception-v3: A machine learning approach for leveling short-range rainfall forecast model from satellite image," in *Advances in Swarm Intelligence*, Y. Tan, Y. Shi, and B. Niu, Eds. Cham: Springer International Publishing, 2019, pp. 105–115.

[11] "California air resources board," 2023. [Online]. Available: https://ww2.arb.ca.gov/resources/documents/air-quality-research-using-satellite-remote-sensing

[12] S. Carnot, "Reflections on the motive power of fire, and on machines fitted to develop that power," *Paris: Bachelier*, vol. 108, no. 1824, 1824.

[13] Y.-H. Chen, T.-J. Yang, J. Emer, and V. Sze, "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 292–308, 2019.

[14] D. A. Crouch and J. Jiru, "Space act agreement 28185 abetween nasa management office for the jet propulsion laboratory and space exploration technologies corp for leo constellation orbit determination and time synchronization support," "Contract agreement", 7 2018.

[15] Y. Cui, C. Ingalz, T. Gao, and A. Heydari, "Total cost of ownership model for datacenter technology evaluation," in *2017 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2017, pp. 936–942.

[16] https://www.darpa.mil/work-with-us/space-based-adaptive-communications-node.

[17] O. L. F. de Carvalho, O. A. de Carvalho Júnior, C. R. e. Silva, A. O. de Albuquerque, N. C. Santana, D. L. Borges, R. A. T. Gomes, and R. F. Guimarães, "Panoptic segmentation meets remote sensing," *Remote Sensing*, vol. 14, no. 4, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/4/965

[18] B. Denby, K. Chintalapudi, R. Chandra, B. Lucia, and S. Noghabi, "Kodan: Addressing the computational bottleneck in space," in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ser. ASPLOS 2023. New York, NY, USA: Association for Computing Machinery, 2023, p. 392–403. [Online]. Available: https://doi.org/10.1145/3582016.3582043

[19] B. Denby and B. Lucia, "Orbital edge computing: Nanosatellite constellations as a new class of computer system," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 939–954. [Online]. Available: https://doi.org/10.1145/3373376.3378473

[20] J. Dexter, J. Sulkin, W. Shimata, J. Dietrick, S. Hnaide, and M. Monson, "We are the spacex software team, ask us anything!" https://www.reddit.com/r/spacex/comments/gxb7j1/we_are_the_spacex_software_team_ask_us_anything/ft6cymf/, 2020.
[21] Z. Dorrani, "Road detection with deep learning in satellite images," *Majlesi Journal of Telecommunication Devices*, vol. 12, no. 1, pp. 43–47, 2023.
[22] N. Drenthe, B. Zandbergen, R. Curran, and M. Van Pelt, "Cost estimating of commercial smallsat launch vehicles," *Acta Astronautica*, vol. 155, pp. 160–169, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0094576518311330
[23] A. R. Duncan, C. M. Szabo, D. P. Bossev, K. A. LaBel, A. M. Williams, M. J. Gadlage, J. D. Ingalls, C. H. Hedge, A. H. Roach, and M. J. Kay, "Single event effects in 14-nm intel microprocessors," in *2016 IEEE Radiation Effects Data Workshop (REDW)*. IEEE, 2016, pp. 1–9.
[24] European-Space-Agency, "Flight ready / modified cots oem camera solutions for space," https://www.raptorphotonics.com/products/space-ready-oem-camera-solutions/, 1 2021.
[25] ——, "Incubed-supported hyperscout imager selected for high-profile australian cubesat mission," https://incubed.esa.int/incubed-supported-hyperscout-imager-selected-for-high-profile-australian-cubesat-mission/, 12 2021.
[26] F. Forcella, R. L. Benech Arnold, R. Sanchez, and C. M. Ghersa, "Modeling seedling emergence," *Field Crops Research*, vol. 67, no. 2, pp. 123–139, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378429000000885
[27] Galorath. (2023) Seer for space. [Online]. Available: https://galorath.com/products/seer-for-space/
[28] R. Giorgiani do Nascimento and F. Viana, "Satellite image classification and segmentation with transfer learning," in *AIAA Scitech 2020 Forum*, 2020, p. 1864.
[29] G. Giuffrida, L. Fanucci, G. Meoni, M. Batič, L. Buckley, A. Dunne, C. Van Dijk, M. Esposito, J. Hefele, N. Vercruyssen *et al.*, "The ϕ-sat-1 mission: The first on-board deep neural network demonstrator for satellite earth observation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.
[30] D. Hardy, M. Kleanthous, I. Sideris, A. G. Saidi, E. Ozer, and Y. Sazeides, "An analytical framework for estimating tco and exploring datacenter design space," in *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2013, pp. 54–63.
[31] A. C. Hax and N. S. Majluf, "Competitive cost dynamics: the experience curve," *Interfaces*, vol. 12, no. 5, pp. 50–61, 1982.
[32] A. M. Hein, R. Matheson, and D. Fries, "A techno-economic analysis of asteroid mining," *Acta Astronautica*, vol. 168, pp. 104–115, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0094576518316357
[33] H. Heiselberg, "Aircraft and ship velocity determination in sentinel-2 multispectral images," *Sensors*, vol. 19, no. 13, p. 2873, 2019.
[34] J. Howard, K. LaBel, M. Carts, R. Stattel, C. Rogers, T. Irwin, and Z. Kahric, "Total ionizing dose testing of the intel pentium iii (p3) and amd k7 microprocessors," Jackson & Tull Chartered Engineers and NASA GSFC and Raytheon ITSS and QSS, Inc and Orbital Sciences Corp, Tech. Rep., 02 2002.
[35] HPE. (2021) HPE Spaceborne Computer. [Online]. Available: https://www.hpe.com/us/en/compute/hpc/supercomputing/spaceborne.html
[36] D. R. Johnson, "Total dose test report on the intel 80486dx2-66 microprocessor tested 8/29-9/8/95," 1995.
[37] H. W. Jones, "Recent large reduction in space launch cost," in *48th International Conference on Environmental Systems*, 2018.
[38] D. Jurakuziev, S. Jumaboev, and M. Lee, "A framework to estimate generating capacities of pv systems using satellite imagery segmentation," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106186, 2023.
[39] L. H. Kaack, G. H. Chen, and M. G. Morgan, "Truck traffic monitoring with satellite images," in *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2019, pp. 155–164.
[40] J. Kang, H. Kim, and Y.-K. Chang, "Development of multiple parameter-based cost model for small earth observation satellite," *26th Annual AIAA/USU Conference on Small Satellites*, 2012.
[41] T. Kerslake and E. Gustafson, "On-orbit performance degradation of the international space station p6 photovoltaic arrays," in *1st International Energy Conversion Engineering Conference (IECEC)*, 2003, p. 5999.
[42] G. Kuperman, T. G. Ulmer, R. Bogoslovov, T. Lockhart, and D. Ott, "Space-based adaptive communications node (Space-BACN)," in *Free-Space Laser Communications XXXIV*, H. Hemmati and B. S. Robinson, Eds., vol. 11993, International Society for Optics and Photonics. SPIE, 2022, p. 119930O. [Online]. Available: https://doi.org/10.1117/12.2614577
[43] K. Kwon, S. Min, J. Kim, and K. Lee, "Framework development for efficient mission-oriented satellite system-level design," *Aerospace*, vol. 10, no. 3, p. 228, 2023.
[44] K. A. LaBel, M. A. Carts, R. A. Gigliuto, C. M. Szabo Jr, M. Kay, T. Sinclair, M. Gadlage, A. Duncan, and D. Ingalls, "Advanced micro devices (amd) processor: radiation test results," in *NEPP Electronics Technology Workshop*, 2013, pp. 11–12.
[45] S. Lee and J. Ahn, "Cost estimation of earth observation satellites based on technology-aware complexity and cost adjustment," *International Journal of Aeronautical and Space Sciences*, vol. 23, no. 2, pp. 379–391, 2022.
[46] Q. Li, S. Wang, X. Ma, A. Zhou, and F. Yang, "Towards sustainable satellite edge computing," in *2021 IEEE International Conference on Edge Computing (EDGE)*, 2021, pp. 1–8.
[47] Y. Li and M. Momen, "Detection of weather events in optical satellite data using deep convolutional neural networks," *Remote Sensing Letters*, vol. 12, no. 12, pp. 1227–1237, 2021.
[48] J. Likar, S. Stone, R. Lombardi, and K. Long, "Novel radiation design approach for cubesat based missions," *24th Annual AIAA/USU Conference on Small Satellites*, 2010.
[49] H. Liu, R. A. Dahlgren, R. E. Larsen, S. M. Devine, L. M. Roche, A. T. O. Geen, A. J. Wong, S. Covello, and Y. Jin, "Estimating rangeland forage production using remote sensing data from a small unmanned aerial system (suas) and planetscope satellite," 1 1970. [Online]. Available: https://agris.fao.org/agris-search/search.do?recordID=CH2023188188
[50] K. Lofstrom, "Server sky — computation and power in orbit," in *2013 1st IEEE Conference on Technologies for Sustainability (SusTech)*, 2013, pp. 126–133.
[51] M. McCord, C. Merry, and P. Goel, "Incorporating satellite imagery in traffic monitoring programs," in *North American Travel Monitoring Exhibitation and Conference*, 1998, pp. 1–18.
[52] P. T. Metzger, "Economics of in-space industry and competitiveness of lunar-derived rocket propellant," *Acta Astronautica*, vol. 207, pp. 425–444, 2023.
[53] F. Millard, "A cloud above the clouds," 2023.
[54] (1996) Working aboard the mir space station. [Online]. Available: \url{https://www.esa.int/esapub/bulletin/bullet88/reite88.htm}
[55] J. R. Moore, "A comparative study of learning curve models in defense airframe cost estimating," Ph.D. dissertation, Air Force Institute of Technology, 2015.
[56] W. Moore, "On the motion of rockets both in nonresisting an resisting mediums," *Journal Natural Phylos. Chem. and Arts*, vol. 27, no. 124, pp. 276–285, 1811.
[57] P. Muthukumar, E. Cocom, K. Nagrecha, D. Comer, I. Burga, J. Taub, C. F. Calvert, J. Holm, and M. Pourhomayoun, "Predicting pm2. 5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data," *Air Quality, Atmosphere & Health*, pp. 1–14, 2021.
[58] Mynaric. (2023) Condor mk3 — optical communication terminal for space-based applications. [Online]. Available: https://mynaric.com/products/space/condor-mk3/
[59] S. Nag, J. LeMoigne, and O. de Weck, "Cost and risk analysis of small satellite constellations for earth observation," in *2014 IEEE Aerospace Conference*, 2014, pp. 1–16.
[60] B. Nagy, J. Farmer, Q. Bui, and J. Trancik, "Statistical basis for predicting technological progress," *PLOS ONE*,

vol. 8, no. 2, pp. 1–7, 02 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0052669

[61] C. Nakalembe, I. Becker-Reshef, R. Bonifacio, G. Hu, M. L. Humber, C. J. Justice, J. Keniston, K. Mwangi, F. Rembold, S. Shukla *et al.*, "A review of satellite-based global agricultural monitoring systems available for africa," *Global food security*, vol. 29, p. 100543, 2021.

[62] "Arset - change detection for land cover mapping," NASA, 2018. [Online]. Available: https://appliedsciences.nasa.gov/join-mission/training/english/arset-change-detection-land-cover-mapping

[63] (2008) New horizons mission powered by space radioisotope power system. [Online]. Available: https://www.energy.gov/ne/articles/new-horizons-mission-powered-space-radioisotope-power-systems

[64] L. Nguyen-Khanh, V. Nguyen-Ngoc-Yen, and H. Dinh-Quoc, "U-net semantic segmentation of digital maps using google satellite images," in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, 2021, pp. 386–391.

[65] Y. Ni, X. Li, Y. Ye, Y. Li, C. Li, and D. Chu, "An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 9, pp. 1545–1549, 2021.

[66] Nippon Telegraph and Telephone. (2022) The key to sustainable growth from space. [Online]. Available: https://www.global.ntt/isf/article-5.html

[67] NVIDIA. (2024) Nvidia h100 tensor core gpu. [Online]. Available: \url{https://www.nvidia.com/en-us/data-center/h100/}

[68] D. of Technical, O. Support, E. Kristiansen, C. Loisy, and W. v. Bosch, *Road Traffic Monitoring by Satellite.* European Space Agency, 2003, vol. 115.

[69] "Aircraft detection," Orbital Insight, 2023. [Online]. Available: https://up42.com/marketplace/blocks/processing/orbital-pleiades-aircraft

[70] P. K. Pandey, S. Jain, and P. Tiwari, "Realization of 1tbps fso/owc based inter satellite link using dp-qpsk for next generation leo satellite internet system," in *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2021, pp. 1–5.

[71] C. Poivey and S. Buchner, "Radiation hardness assurance (rha) for space systems," in *SERESSA-3rd International School on the Effects of Radiation on Embedded Systems for Space Applications*, 2007.

[72] W. Powell, M. Campola, T. Sheets, A. Davidson, and S. Welsh, "Commercial off-the-shelf gpu qualification for space applications," NASA, Space Technology Mission Directorate Annual Program Review, Cleveland OH, USA, Tech. Rep., 9 2018.

[73] A. S. Rakin, Z. He, and D. Fan, "Bit-flip attack: Crushing neural network with progressive bit search," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1211–1220.

[74] K. Sahu, "Radiation report on trmm/gpep part no. mq80386-20, ppm-93-062," 1993.

[75] P. K. Sarangi, B. Sharma, L. Rani, and M. Dutta, "Satellite image classification using convolutional neural network," *Advances in Aerial Sensing and Imaging*, pp. 333–353, 2024.

[76] C. Schorn, A. Guntoro, and G. Ascheid, "An efficient bit-flip resilience optimization method for deep neural networks," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2019, pp. 1507–1512.

[77] S. Singha, T. J. Bellerby, and O. Trieschmann, "Satellite oil spill detection using artificial neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 6, pp. 2355–2363, 2013.

[78] SpaceX, "Sustainability," https://www.spacex.com/updates/ /#sustainability, 02 2022.

[79] C. M. Szabo, A. Duncan, K. A. LaBel, M. Kay, P. Bruner, M. Krzesniak, and L. Dong, "Preliminary radiation testing of a state-of-the-art commercial 14nm cmos processor/system-on-a-chip," in *2015 IEEE Radiation Effects Data Workshop (REDW)*. IEEE, 2015, pp. 1–8.

[80] S. Tehsin, S. Kausar, A. Jameel, M. Humayun, and D. K. Almofarreh, "Satellite image categorization using scalable deep learning," *Applied Sciences*, vol. 13, no. 8, p. 5108, 2023.

[81] The-Boeing-Company, https://www.nasa.gov/wp-content/uploads/2021/02/473486main_iss_atcs_overview.pdf.

[82] K. Thome, "Satellite observations," 2007. [Online]. Available: https://terra.nasa.gov/citizen-science/air-quality/part-ii-track-pollution-from-space

[83] Z. Tong, Y. Li, Y. Li, K. Fan, Y. Si, and L. He, "New network based on unet++ and densenet for building extraction from high resolution satellite imagery," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 2268–2271.

[84] G. M. Ung, "Finally, a good use for intel's lowly atom cpu: Inside nasa's $2.7b perseverance rover," https://www.pcworld.com/article/394482/the-nasa-perseverance-rover-uses-intels-atom-cpu.html, 4 2021.

[85] K. Vani *et al.*, "Deep learning based forest fire classification and detection in satellite images," in *2019 11th International Conference on Advanced Computing (ICoAC)*. IEEE, 2019, pp. 61–65.

[86] D. Vasisht, J. Shenoy, and R. Chandra, "L2d2: low latency distributed downlink for leo satellites," in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, ser. SIGCOMM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 151–164. [Online]. Available: https://doi.org/10.1145/3452296.3472932

[87] S. Voigt, F. Giulio-Tonolo, J. Lyons, J. Kučera, B. Jones, T. Schneiderhan, G. Platzeck, K. Kaku, M. K. Hazarika, L. Czaran *et al.*, "Global trends in satellite-based emergency mapping," *Science*, vol. 353, no. 6296, pp. 247–252, 2016.

[88] M. Wang, H. Qiu, L. Xu, D. Wang, Y. Li, T. Zhang, J. Liu, and H. Li, "A case for application-aware space radiation tolerance in orbital computing," *arXiv preprint arXiv:2407.11853*, 2024.

[89] M. Wieland and S. Martinis, "A modular processing chain for automated flood monitoring from multi-spectral satellite data," *Remote Sensing*, vol. 11, no. 19, p. 2330, 2019.

[90] M. Wieland, S. Martinis, R. Kiefl, and V. Gstaiger, "Semantic segmentation of water bodies in very high-resolution satellite and aerial images," *Remote Sensing of Environment*, vol. 287, p. 113452, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425723000032

[91] M. Wijnen, N. Agüera-Lopez, S. Correyero-Plaza, and D. Perez-Grande, "Cubesat lunar positioning system enabled by novel on-board electric propulsion," *IEEE Transactions on Plasma Science*, vol. 46, no. 2, pp. 319–329, 2018.

[92] D. Woolf. (2023) Low-alpha, variable emissivity radiator (laver) panels for passive thermal regulation of spacecraft. [Online]. Available: \url{https://tfaws.nasa.gov/wp-content/uploads/TFAWS23-PT-25-Presentation.pdf#page=9}

[93] T. P. WRIGHT, "Factors affecting the cost of airplanes," *Journal of the Aeronautical Sciences*, vol. 3, no. 4, pp. 122–128, 1936.

[94] H. Wu, H. Zhang, J. Zhang, and F. Xu, "Fast aircraft detection in satellite images based on convolutional neural networks," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 4210–4214.

[95] Y. N. Wu, P.-A. Tsai, A. Parashar, V. Sze, and J. S. Emer, "Sparseloop: An analytical approach to sparse tensor accelerator modeling," in *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2022, pp. 1377–1395.

[96] Y. Xiao, J. Liu, Y. Shen, X. Jiang, and N. Shiratori, "Secure communication in non-geostationary orbit satellite systems: A physical layer security perspective," *IEEE Access*, vol. 7, pp. 3371–3382, 2019.

[97] S. Yang, 2007. [Online]. Available: http://www.stats.gov.cn/english/icas/papers/p020071017422083591595.pdf

[98] C. Yao, W. Liu, W. Tang, and S. Hu, "Eais: Energy-aware adaptive scheduling for cnn inference on high-performance gpus," *Future Generation Computer Systems*, vol. 130, pp. 253–268, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X22000127

[99] Y. Zhang, J. Ye, G. Pan, and M.-S. Alouini, "Secrecy outage analysis for satellite-terrestrial downlink transmissions," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1643–1647, 2020.