A Satellite-Born Server Design with Massive Tiny Chips Towards In-Space Computing

Mengwei Xu, Li Zhang, Hongyu Li, Ruolin Xing, Qibo Sun
Beijing University of Posts and Telecommunications

With the explosive data volume generated by satellites, in-space computing becomes indispensable to fully exploit the value of such data. Similar to today's computing paradigms on the ground, a key step to the success of in-space computing is building distributed, micro data centers in space, which consist of many server machines. In this work, we propose a novel server architecture that is comprised of massive, low-power system-on-chips (SoCs). Through quantitative analysis, we confirm that such a server significantly outperforms conventional servers (Intel CPU, NVIDIA GPU, etc) on three critical metrics of in-space computing, i.e., energy efficiency, weight, and volume. Notably, we also demonstrate that the three concerned metrics above can be reduced to only one (energy efficiency) as the accompanied solar panels to provide the needed energy often overwhelm the server itself in weight and volume. We further dive into two specific applications that are representative of in-space computing: video processing and deep learning inference. Through state-of-the-art software and benchmarks, we reveal that our SoC-based server, especially with its heterogeneous processors (GPU, DSP, and MediaCodec), can remarkably improve the energy efficiency compared to conventional servers.

I. Introduction

Comprised of massive low Earth orbit (LEO) satellites, emerging constellations [1] are capable of delivering almost real-time high-resolution Earth imagery. Yet, transmitting all of such data to the ground is impractical due to the constrained space-ground network link, nor it is necessary as many of the images are of clouds or empty seas. The straightforward solution is (pre-)processing those data in orbit and only getting the critical/valuable portion to the ground. Beyond Earth imagery, there are many other kinds of data captured in space and are better to be processed in space before getting transmitted to the ground as well, e.g., the scientific experiment results, satellites operational statistics, and astronauts' health status [2].

The above applications and scenarios call for **in-space computing**. How to efficiently deliver computing resources above sky? A lesson we have learned from the many-years evolution of computing on the ground is that: instead of equipping each device with adequate computing capacity, it is much more economical to build a centralized powerful datacenter where weak devices can offload their tasks using a C/S paradigm. Correspondingly, we envision that there will be many micro datacenters on distributed satellites that provide computing and storage resources to nearby satellites that do not have

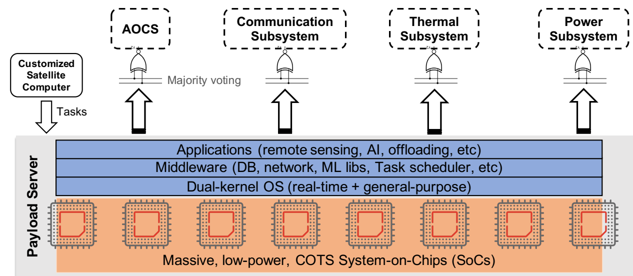

Fig. 1. The proposed satellite architecture for in-space computing.

adequate computing resource, e.g., nanosatellites. Such in-space datacenters are becoming the frontier for computing in the next decades. Major cloud providers on the ground such as AWS and Azure are attempting to extend their services with satellites [3], [4]. Meanwhile, startups like OrbitsEdge and Loft Orbital have launched commercial-off-the-shelf (COTS) rack servers into the space to promote the concept of "space infrastructure as a service" [5]. As an initial step, it is of primary importance to design the in-space server machines – the basic unit that constitutes a data center.

What makes a good server machine in space? Notably, designing in-space datacenters is extraordinarily different from the ones on the ground. The first contribution of this work is the proposal of a set of criteria to quantify the efficiency of a server in space. At the rocket launching time, the killing factors of the server are its weight and volume. Therefore, we use throughput per weight (TpW) and throughput per volume (TpV) as two critical metrics. When a server is operating in space, the energy harvested through a solar panel becomes the primary constraint. Therefore, we use throughput per energy (TpE). The throughput used in the above metrics indicates how much computing capacity is provided and relates to specific applications or workloads, e.g., the number of Earth images processed per second. Reliability is another killing factor towards in-space datacenter due to its harsh operational environment. To this end, hardware hardening and redundancy techniques have been widely adopted [6].

The conventional servers are not designed or optimized for the above criteria, and therefore are not likely to be adequate for operation in space. In this work, we propose a new form of servers for space, namely SoC-Cluster, which consists of tens or hundreds of mobile SoCs like Qualcomm Snapdragon or Apple Silicon. As shown in Figure 1, one or multiple SoC-Clusters can be carried in one satellite and provide high computing capacity for the satellite itself or the tasks offloaded from other satellites or the ground. The underlying rationale is that mobile SoCs are designed to be tiny, lightweight, and energy-efficient by using a simpler instruction set (mostly ARM) and smaller transistors than traditional high-end processors. For instance, the Qualcomm Snapdragon 888 SoC released in 2021 is based on 5nm technology while Xeon is still using $\geq$10nm. Moreover, each SoC can be independently managed (e.g., turned on/off and frequency control) to adapt to the dynamic workloads, which is much more flexible than a monolithic powerful processor like NVIDIA GPU.

SoC-Cluster can also potentially improve satellite reliability. Unlike traditional satellites that rely on hardware hardening techniques [6], SoC-Cluster can leverage its massive SoCs to enhance reliability through *voting*. As shown in Figure 1, each subsystem is connected to many or all SoCs; between them, there is a voting circuit that takes output signals from the SoCs and obtains only the majority to the subsystem. By controlling how many SoCs participate in the decision-making, we can easily trade off computing/energy redundancy and reliability for each subsystem or task. Moreover, SoC-Cluster is highly robust under harsh in-space environment: one or some of the SoCs being destroyed by the ionizing particles does not affect the whole system functionality as long as there are enough SoCs still operating normally. In comparison, traditional monolithic servers can hardly recover from severe hardware failure.

SoC-Cluster is not a pie-in-the-sky vision. Instead, it has been densely deployed by one of the largest edge service providers, Alibaba Edge Node Service (ENS), on ground edges. Those SoC-Clusters are mainly supporting one specific application, i.e., cloud gaming [7]–[10], which enables wimpy or low-battery smartphones to run resource-consuming games anytime and anywhere. §II-B will give a detailed description of how such SoC-Cluster is constructed and deployed. However, the potential of those SoC-Clusters is far from being fully realized. It is unsure if those servers can efficiently serve typical in-space workloads.

To understand whether SoC-Cluster can efficiently operate in space, we perform a first-of-its-kind measurement of typical servers. The measurement consists of two major parts. First, we analyze the theoretical TpW, TpV, and TpE of typical servers, including Intel Xeon CPU, NVIDIA GPU, PowerEdge machines with multiple CPU configurations, etc. We summarize the results into two major observations. (1) SoC-Cluster has significantly higher throughput under the constraints of weight, volume, and energy than other conventional servers. (2) Among the three metrics proposed, energy (TpE) shall be the primary concern. This is because harvesting energy requires solar panels which take extra weight and volume to be launched by rockets. The weight and volume of those solar panels are much larger than the server itself to fully utilize its computing capacity. In other words, the three constraints can be reduced to only one. This finding is valuable in simplifying the server design for space.

The second part of our measurement is a workload-driven study. More specifically, we quantify the killing metric TpE of SoC-Cluster and conventional servers using two concrete workloads: video processing and deep learning inference. Those two workloads promise to be the key building block for representative in-space applications such as remote sensing. For instance, a sequence of Earth images captured by a satellite might need to be transcoded first from high to low resolution and then fed to a deep learning model for prediction. To this end, we build an automatic benchmark suite that leverages state-of-the-art libraries for each tested workload. Our key observation is that SoC-Cluster delivers significantly higher energy efficiency than conventional servers. For instance, SoC-Cluster is able to transcode 67–93 1080p frames per Joule, which is 9.3–15.5$\times$ higher than Intel CPU and about 10$\times$ higher than NVIDIA GPU that we used in experiment.

The major contributions of this work are summarized as follows.

- We identify a few key metrics for efficient in-space computing. Through quantitative analysis, we further narrow them down to simplify the design and testing of in-space datacenter servers (i.e., energy efficiency covers the metrics of volume and weight).
- We propose a novel satellite-born server architecture, namely SoC-Cluster, that consists of massive, low-power SoCs. On the metrics proposed above, SoC-Cluster significantly outperforms conventional servers such as Intel CPU and NVIDIA GPU.
- We perform concrete, comparative experiments on SoC-Cluster and conventional ones with two applications that are representative of in-space computing. The results highlight the superior energy efficiency of SoC-Cluster.

II. Rationales and Related Work

A. *Computing in Space*

Space is the new frontier for computing. There are a few reasons to push computations from the ground to space.

• **Space-native data.** Satellites or international space stations (ISS) are generating massive data such as Earth imagery, weather observations, and cosmic rays statistics. That data is not likely to be all streamed to the ground due to the constraint of the space-ground network link and the high cost of ground station rental. Instead, the data needs to be (pre-)processed, and only critical or more compact results should be transmitted to the ground. For example, a small NN model can be deployed on satellites to filter out the images containing no objects of interest and only the valuable Earth images will be sent to the ground for further analysis [11].

• **High availability.** Nowadays, the majority of the Earth and billions of people still have no access to the Internet. While satellite constellations like Starlink promise to deliver the broadband Internet service to those areas, certain applications cannot tolerate a long network routing to a remote cloud data center. For example, computing in orbit could benefit a multi-user gaming application. In-space datacenters could be an essential extension of edge computing and the "near-data processing" paradigm.

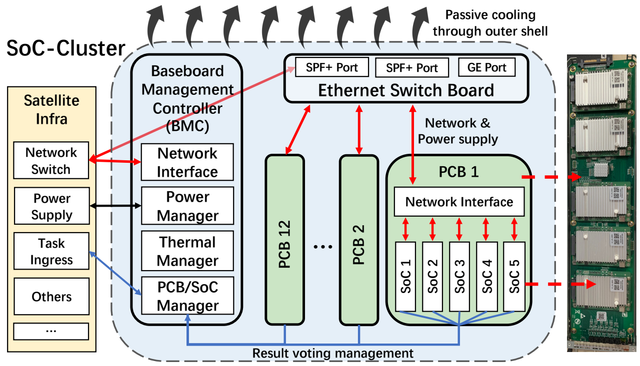

Fig. 2. Architecture and network topology of the SoC server used in our experiment.

• **Green energy.** Satellites are born with zero-carbon nature. Compared to the solar energy harvested on the ground, satellites have a much higher solar-to-electricity converting ratio without the Earth's atmosphere. Deploying a data center in space takes an important step toward green computing.

What matters to the data centers or servers in space? To deploy a server in space, there are two major steps. First, it needs to be launched to a specified orbit using a rocket. At this stage, weight and volume are the two major constraints. Second, once a server is operating in space, the amount of energy harvested and runtime reliability become crucial factors. In this work, we mainly focus on quantitative analysis and measurement of weight, volume, and energy with each respective metric. Reliability has been extensively studied in relevant literature [12]. Our SoC-Cluster inherently provides reliability benefit through a decentralized architecture design, therefore is not a focus of this work.

B. SoC-Cluster Overview

The design space is huge in turning the idea of SoC-Cluster into a real edge server. To carry out concrete experiments, we use one SoC-Cluster server from a leading manufacturer in China. This type of SoC-Cluster has been densely deployed in the wild by Alibaba ENS, therefore is representative of the status quo of SoC-Cluster. For simplicity, in the rest of the paper, we still refer to this particular server as SoC-Cluster. Fig. 2 shows the architecture of the SoC-Cluster used in our measurements. The major component of the server is a pool of 60 Qualcomm Snapdragon 865 chips [13]. Each SoC contains an Octa-core CPU, an Adreno 650 GPU, a Hexagon 695 digital signal processor (DSP), and a 12GB LPDDR5 DRAM. The server is organized as 12 PCBs where each PCB integrates 5 SoCs. The PCBs in SoC-Cluster provide both power supply and network capabilities (served as a switch) for SoCs attached to them. Network interface bandwidth between SoC and PCB is constrained to 1 Gbps. Plugging PCB to SoC-Cluster will establish a physical connection with the Ethernet Switch Board (ESB) in SoC-Cluster. The network capacity between each PCB and ESB is also 1 Gbps. The ESB is responsible for exposing all SoCs for external access through its single RJ45 interface (1 Gbps) or dual SPF+ interfaces (2×10Gbps). In our case, we use an external network switch (10Gbps) to connect the external machine and SoC-Cluster through their SPF+ interfaces to ensure our benchmarks will not be bounded by network throughput. SoC-Cluster also uses a Baseboard Management Controller (BMC) to manage and monitor the server status (e.g., power, temperature, and fans).

Alibaba ENS, a leading edge service provider worldwide, has deployed thousands of SoC-Cluster machines on their edge sites for almost two years. Currently, those SoC-Clusters are serving only one application: cloud gaming, as the Android-native games like Genshin Impact, can be seamlessly supported by the mobile SoC without any revision. Cloud gaming deployed on edges enables wimpy or low-power devices to run resource-hungry games with a good user experience. Developers are now able to access the resources of those SoC-Clusters through either a virtualized or native context just like traditional cloud services. In total, Alibaba ENS serves millions of game sessions per day.

We are not the first to conceptualize a server consisting of tiny SoCs. There are attempts [14], [15] to investigate whether mobile SoCs can provide sufficient performance and reduce costs for HPC. To reduce e-waste, Shahrad et al. [16] build computation nodes with used smartphones and gave an analysis of server design, but didn't evaluate real workloads. Switzer et al. use only five smartphones to build a junkyard data center [17] with carbon concerns. Some work uses IoT/mobile SoCs to support specific applications, like video transcoding [18], key-value storage [19], and parallel computing [20]. Those work mostly focus on specific apps and lack performance comparison with conventional servers.

Different from the above work, we envision SoC-Cluster to be the building block for in-space data centers. We use a COTS SoC-Cluster and perform an application-driven measurement on it in the context of space, e.g., using metrics and workloads representative of the space scenario. Apart from the vantages in size, weight, energy efficiency, and reliability as discussed in §I, a few more incentives lead us to explore SoC-Cluster in space. First, mobile SoCs are highly scalable: each SoC has a standard CPU that can serve general workloads, but also heterogeneous co-processors like GPU, DSP, and NPU [21] that accelerate domain-specific workloads. Mobile SoC is still fast evolving [22], which congregates the wisdom of many leading chip companies. Building a server atop those SoCs takes such free lunch. Furthermore, the software stacks on mobile SoCs and OSes are mature enough and highly optimized, e.g., deep learning inference/training [23], [24], multimedia data processing [25], or even containers [26], [27].

III. Quantitative Analysis

In this section, we quantify the performance of SoC-Cluster and conventional servers using the three metrics we propose in this work: throughput per weight (TpW), throughput per volume (TpV), and throughput per energy (TpE). The throughput here is represented by the theoretical computing capacity, i.e., floating point operations per second (FLOPs) and integer

	Throughput per Energy (TpE)			Throughput per Volume (TpV)			Throughput per Weight (TpW)		
	Power (watt)	**GFLOPs per watt (FP32)**	**GINOPs per watt (INT8)**	**Volume (U)**	**GFLOPs per U (FP32)**	**GINOPs per U (INT8)**	**Weight (kg)**	**GFLOPs per kg (FP32)**	**GINOPs per kg (INT8)**
Xeon 40-core CPU Server	276.3	0.8	0.5	1	208.3	130.4	18.8	11.1	6.9
NVIDIA A40 GPU Server	2,000.0	149.6	1,197.2	4	74,800	598,600	57.9	5,165.8	41,339.8
PowerEdge R350	95.0	0.5	0.9	1	49.3	85.4	13.6	3.6	6.3
PowerEdge R550	330.0	0.5	0.9	2	83.0	151.3	20.4	8.1	14.8
PowerEdge R750xs	370.0	0.6	1.0	2	104.6	182.5	21.9	9.5	16.6
SoC-Cluster (Kryo CPU)	672.0	1.3	0.2	2	437.4	76.5	27.0	32.4	5.7
SoC-Cluster (Adreno GPU)	387.0	193.8	X	2	37,500	X	27.0	2,777.8	X
SoC-Cluster (Hexagon DSP)	345.5	X	2,604.9	2	X	450,000	27.0	X	33,333.3

TABLE I

THEORETICAL COMPARISON BETWEEN SOC-CLUSTER AND CONVENTIONAL COTS EDGE SERVERS. "X" MEANS THAT THIS NUMERICAL OPERATION IS NOT SUPPORTED BY THE HARDWARE.

	Server Volume	Solar Panel Volume	Server Weight	Solar Panel Weight
Xeon 40-core CPU server	1	1.7–4.3	18.8	27.6–120.1
NVIDIA A40 GPU server	4	121.2–312.5	57.9	2,000.0–8,695.7
SoC-Cluster	2	4.1–10.5	27.0	67.2–292.2

TABLE II

BOTTLENECK ANALYSIS OF IN-SPACE COMPUTING: THE SOLAR PANELS DEMANDED TO PROVIDE ENOUGH POWER IS MUCH HEAVIER AND LARGER THAN THE SERVER ITSELF. WE ASSUME THE AVERAGE SERVER UTILIZATION IS 50%.

operations per second (INOPs). We compare SoC-Cluster to the most representative edge servers: Intel CPU server, NVIDIA GPU server, and PowerEdge series are equipped with different series of Intel CPU. We obtain their operation numbers (capacity, power, volume, and weight) through the public datasheets found on the Internet.

Performance comparison. We summarize our analytic results in Table I. Our key observation is that SoC-Cluster has significantly higher TpE, TpW, and TpV than conventional servers. For instance, the TpE of SoC-Cluster of FP32 operations is 1.7×/2.5× higher than the Xeon CPU server and PowerEdge R350, respectively. The SoC Adreno GPU is even 2 orders of magnitudes higher in power efficiency. SoC GPU is also more power efficient than NVIDIA GPUs (193.8 vs. 149.6 GFLOPs per watt). For 8-bit integer operations (INT8), the SoC digital signal processor (DSP) achieves even higher benefits in TpE, e.g., 2.2× higher than NVIDIA A40 GPU and 3 orders of magnitude higher CPU servers. This is mainly because embedded DSPs are specifically designed for low-power scenarios.

NVIDIA A40 GPU outperforms SoC-Cluster in TpW and TpV. This is mainly because modern datacenter-level GPUs are highly optimized with a tremendous number of CUDA cores and thus mighty parallelism capacity. However, as we will show next, among the three metrics proposed, TpE is often the bottleneck metric for in-space computing and thus SoC-Cluster is still the best option.

Energy bottleneck analysis. We then dive into the three metrics and seek to identify the most killing factor among them: which of them is more likely to be the bottleneck? The key rationale is that in-space servers are sustained by solar panels that harvest energy from solar power. Therefore, the solar panel adds extra weight and volume to the satellite at launch time, which shall be jointly considered with the server itself. To understand the relative ratio of such indirect overhead to the physical server, we look into the statistics of popular solar panel products [28], especially the energy that can be harvested per weight (E_w, unit: Watts/kg) and per volume (E_v, unit: Watts/U). We conclude that the state-of-the-art solar panels used in small spacecraft can provide 64–165 Watts/kg or 2.3–10 Watts/U power supply. Assuming that the solar panels can harvest solar energy half of the time (at other times it will be blocked by the earth) and the server's peak energy consumption is E_{peak}, we can estimate the weight and volume of the solar panels demanded to keep the servers operate at a utilization α.

$$Weight_{panel} = \alpha * E_{peak} * 2/E_w$$

$$Volume_{panel} = \alpha * E_{peak} * 2/E_v$$

Based on the above analytical model, we summarize the weight/volume of the server and the solar panels demanded to provide 50% peak utilization (α) in Table II. The results show that to provide the power for 50% hardware utilization, the solar panels are much larger/heavier than the server itself. For example, to support the operation of the NVIDIA A40 (x8) GPU server, the minimal volume and weight of the qualified solar panel is more than 100U and 2,000kg, which is about 30×–35× more than the GPU server itself. In other words, even if the GPU server can provide high computing capacity, in a real situation only a small portion of its power can be released. To be noted, the solar panels typically occupy less than 50% of the whole satellite's weight and size. In that case, the metrics of TpW and TpV are covered by the energy constraint.

In summary, as constrained by the SOTA energy harvesting technique, energy efficiency is the primary limiting factor

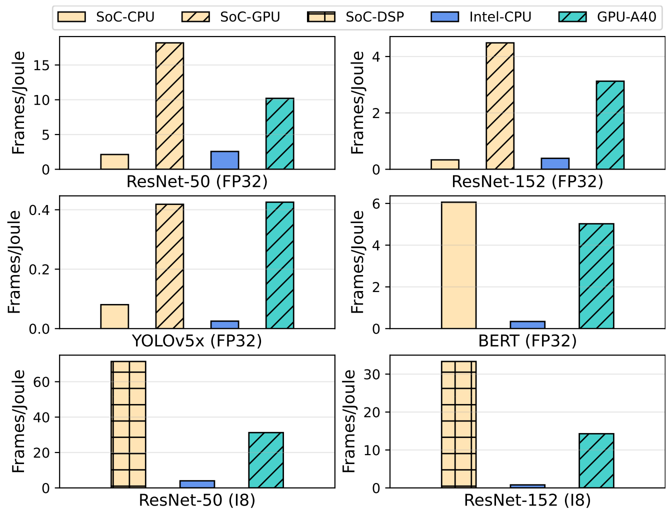

Fig. 3. Processed frames per Joule (an indicator of TpE) on the deep learning inference experiments. FP32: 32-bit floating point; I8: 8-bit integer.

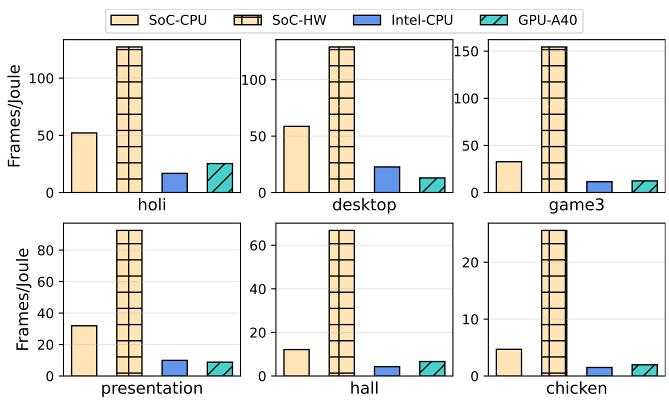

Fig. 4. Processed frames of per Joule (an indicator of TpE) on the video processing experiments. The 6 videos are randomly selected from a popular video benchmark [29].

(as against weight and volume) of in-space computing. Such insight can greatly simplify the design and testing of satellite-borne data centers or servers in the future.

IV. Workloads-driven Experiments

In this section, we further quantitatively investigate the performance of SoC-Cluster with two specific applications, video processing (transcoding) and deep learning inference. Those two applications are typical for in-space computing, especially remote sensing: (i) the Earth imagery is often ingested as videos on satellites and needs to be transcoded before either being sent to ground stations or processed directly. Such a pre-processing step is dispensable as the sensing data generated by the satellites could be overwhelmingly large as against to the constrained space-ground network capacity. (ii) the Earth imagery is often processed using DL models to detect the objects (vehicles, ships, etc.) or filter out the low-quality data (e.g., those covered by clouds).

Using them as case studies, we want to answer the critical question about how efficiently SoC-Cluster can serve in-space workloads as compared to conventional servers. Specifically, we use an edge-typical server with a 40-core (80-thread) Intel Xeon Gold 5218R processor and 8 NVIDIA A40 GPUs (released in the same year as Snapdragon 865) for comparison. We only focus on the energy efficiency of those servers as we have shown that it is the most concerned metric for in-space computing.

Software. For video transcoding, we use FFmpeg (v4.4) [30] with H.264 codec support. We cross-compile FFmpeg to SoC-Cluster with ARMv8 NEON acceleration. To utilize the hardware codec of Qualcomm SoCs, we use a popular open-source Android library LiTr [31]. We randomly pick 6 videos from vbench [29], a widely used benchmark tool for cloud video transcoding. For DL serving, we use ResNet-50, ResNet-152, YOLOv5x, and BERT for CV and NLP tasks. Those models are extensively used in CV/NLP tasks. We experiment using TFLite [24] and MNN [23] on ARM SoC, TVM [32] on Intel CPU, and TensorRT [33] on NVIDIA GPU, considering their popularity and state-of-the-art performance. The power consumption of SoC-Cluster is measured through the software-level APIs exposed through its BMC, which includes not only the power of SoCs but also the PCBs and fans. The workload energy consumption report is subtracted by the idle power consumption.

Results. The results of DL inference and video transcoding are illustrated in Figure 3 and Figure 4, respectively. The Y-axis shows the energy efficiency, represented by the frames processed per Joule. Our key observation is that, *SoC-Cluster is significantly more energy-efficiency than conventional CPU/GPU servers for both two applications.* Recall that the analysis in §III indicates that energy efficiency is the killing factor to in-space computing, we can conclude that SoC-Cluster is able to handle more workloads than conventional servers, even though their theoretical maximal computing capacity is not that high.

- **DL inference energy efficiency.** As illustrated in Figure 3, our key observation is that SoC-Cluster, especially its domain-specific accelerators like GPU and DSP, provides significantly higher energy efficiency than conventional servers. Running prediction with ResNet-50 model (FP32), SoC GPU can process 18.2 samples per Joule, which is 7× and 1.8× higher than Intel CPU and NVIDIA GPU, respectively. The energy efficiency of SoC DSP is even more significant, i.e., 2.3× higher than NVIDIA A40 GPU (with batch size 64).

SoC-Cluster not only delivers higher energy efficiency but can also proportionally scales such efficiency with the workloads. When the workload is lightweight (i.e., using a small batch size), the energy efficiency of NVIDIA GPU further drops, e.g., 10.2 to 2.8 samples per Joule with ResNet-50 (FP32). Instead, SoCs can process each sample efficiently when batch size is one, and some of them can be turned off to eliminate the energy waste without workloads. We believe that such a feature will be highly valued in space as the workloads are expected to be variational due to satellites' continuous, high-speed movement.

- **Video processing energy efficiency.** SoC-Cluster's advantage in energy efficiency is even more significant compared to conventional servers in the video processing applications. SoC-Cluster's hardware codec can transcode 26–154 frames per Joule, which is 5.7×–17.1× higher than Intel CPU and 5.0×–13.0× higher than NVIDIA A40 GPU. Even without using the hardware codec, SoC CPU is still a few times more energy-efficient than conventional servers. Our additional experiments also confirm that the quality of the transcoded videos is consistent with the ones by conventional servers.

V. Conclusion

In this work, we propose a new server architecture called SoC-Cluster for efficient in-space computing. It consists of numerous low-power SoCs with high computing capability and low power consumption. We examine the SoC-Cluster in two parts. First, we analyze the performance of different servers from many aspects related to the launching and operating phases of a satellite. The results highlight the superior performance of SoC-Cluster. Second, we compare SoC-Cluster and traditional servers atop two typical in-space workloads: video processing and deep learning inference. The end-to-end application experiments demonstrate that SoC-Cluster is much more energy-efficient than conventional servers.

Acknowledgment

This work was supported by National Key R&D Program of China (No.2021ZD0113001), Beijing Nova Program (No.Z211100002121118), and Young Elite Scientists Sponsorship Program by CAST (No.2021QNRC001).

References

[1] S. Wang, Q. Li, M. Xu, X. Ma, A. Zhou, and Q. Sun, "Tiansuan Constellation: An Open Research Platform," *arXiv preprint arXiv:2112.14957*, 2021.

[2] "International Space Station Benefits for Humanity," https://www.nasa.gov/stationbenefits, 2022.

[3] "The Space Cloud: Satellite Strategies for AWS, Google and Microsoft," https://datacenterfrontier.com/the-space-cloud-satellite-strategies-for-aws-google-and-microsoft/, 2021.

[4] "Azure Space," https://azure.microsoft.com/en-us/solutions/space/#overview, 2022.

[5] "Data Centers Above the Clouds: Colocation Goes to Space," https://datacenterfrontier.com/data-centers-above-the-clouds-colocation-goes-to-space/, 2020.

[6] A. Keys, J. Adams, J. Cressler, M. Johnson, and M. Patrick, "A review of NASA's radiation-hardened electronics for space environments project," in *AIAA SPACE 2008 Conference & Exposition*, 2008, p. 7673.

[7] "Amazon Luna," https://www.amazon.com/luna/landing-page, 2022.

[8] "GeForce Now," https://www.nvidia.com/en-us/geforce-now/, 2022.

[9] "Google Stadia," https://stadia.google.com/, 2022.

[10] "Cloud Gaming, Meet Facebook Gaming," https://www.facebook.com/fbgaminghome/blog/cloud-gaming-meetfacebook-gaming, 2022.

[11] "Hewlett Packard Enterprise drives innovation at the extreme edge on the International Space Station with 24 completed experiments," https://www.hpe.com/us/en/newsroom/press-release/2022/04/hewlett-packard-enterprise-drives-innovation-at-the-extreme-edge-on-the-international-space-station-with-24-completed-experiments.html, 2022.

[12] J. Budroweit and H. Patscheider, "Risk assessment for the use of cots devices in space systems under consideration of radiation effects," *Electronics*, vol. 10, no. 9, p. 1008, 2021.

[13] "Hardware — Qualcomm Snapdragon 865 5G Mobile Platform — 5G Mobile Processor — Qualcomm," https://www.qualcomm.com/products/application/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-865-5g-mobile-platform, 2020.

[14] N. Rajovic, P. M. Carpenter, I. Gelado, N. Puzovic, A. Ramirez, and M. Valero, "Supercomputing with commodity CPUs: Are mobile SoCs ready for HPC?" in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 2013, pp. 1–12.

[15] N. Rajovic, A. Rico, N. Puzovic, C. Adeniyi-Jones, and A. Ramirez, "Tibidabo: Making the case for an ARM-based HPC system," *Future Generation Computer Systems*, vol. 36, pp. 322–334, 2014.

[16] M. Shahrad and D. Wentzlaff, "Towards Deploying Decommissioned Mobile Devices as Cheap {Energy-Efficient} Compute Nodes," in *9th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 17)*, 2017.

[17] J. Switzer, R. Kastner, and P. Pannuto, "Architecture of a Junkyard Datacenter," *arXiv preprint arXiv:2110.06870*, 2021.

[18] P. Liu, J. Yoon, L. Johnson, and S. Banerjee, "Greening the Video Transcoding Service with {Low-Cost} Hardware Transcoders," in *2016 USENIX Annual Technical Conference (USENIX ATC 16)*, 2016, pp. 407–419.

[19] D. G. Andersen, J. Franklin, M. Kaminsky, A. Phanishayee, L. Tan, and V. Vasudevan, "FAWN: A fast array of wimpy nodes," in *Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles*, 2009, pp. 1–14.

[20] F. Büsching, S. Schildt, and L. Wolf, "Droidcluster: Towards smartphone cluster computing–the streets are paved with potential computer clusters," in *2012 32nd International Conference on Distributed Computing Systems Workshops*. IEEE, 2012, pp. 114–117.

[21] "Kirin 990 5G chipset," https://www.hisilicon.com/en/products/Kirin/Kirin-flagship-chips/Kirin-990-5G, 2019.

[22] "Qualcomm Showcases Future Technology Roadmap to Drive the Connected Intelligent Edge and Lead the World to 5G Advanced and Beyond," https://www.qualcomm.com/news/releases/2022/02/28/qualcomm-showcases-future-technology-roadmap-drive-connected-intelligent, 2022.

[23] X. Jiang, H. Wang, Y. Chen, Z. Wu, L. Wang, B. Zou, Y. Yang, Z. Cui, Y. Cai, T. Yu *et al.*, "MNN: A universal and efficient inference engine," *arXiv preprint arXiv:2002.12418*, 2020.

[24] "TensorFlow Lite," https://www.tensorflow.org/lite, 2022.

[25] "MediaCodec," https://developer.android.com/reference/android/media/MediaCodec, 2022.

[26] W. Song, J. Ming, L. Jiang, Y. Xiang, X. Pan, J. Fu, and G. Peng, "Towards Transparent and Stealthy Android OS Sandboxing via Customizable Container-Based Virtualization," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 28582874. [Online]. Available: https://doi.org/10.1145/3460120.3484544

[27] W. Chen, L. Xu, G. Li, and Y. Xiang, "A lightweight virtualization solution for android devices," *IEEE Transactions on Computers*, vol. 64, no. 10, pp. 2741–2751, 2015.

[28] "State-of-the-art of small spacecraft technology," https://www.nasa.gov/smallsat-institute/sst-soa/power, 2021.

[29] A. Lottarini, A. Ramírez, J. Coburn, M. A. Kim, P. Ranganathan, D. Stodolsky, and M. Wachsler, "vbench: Benchmarking Video Transcoding in the Cloud," in *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS 2018, Williamsburg, VA, USA, March 24-28, 2018*, X. Shen, J. Tuck, R. Bianchini, and V. Sarkar, Eds. ACM, 2018, pp. 797–809. [Online]. Available: https://doi.org/10.1145/3173162.3173207

[30] "FFmpeg," https://ffmpeg.org/, 2022.

[31] "linkedin/LiTr: Lightweight hardware accelerated video/audio transcoder for Android." https://github.com/linkedin/LiTr, 2022.

[32] T. Chen, T. Moreau, Z. Jiang, L. Zheng, E. Yan, H. Shen, M. Cowan, L. Wang, Y. Hu, L. Ceze, C. Guestrin, and A. Krishnamurthy, "TVM: An Automated End-to-End Optimizing Compiler for Deep Learning," in *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*. Carlsbad, CA: USENIX Association, Oct. 2018, pp. 578–594. [Online]. Available: https://www.usenix.org/conference/osdi18/presentation/chen

[33] "NVIDIA TensorRT," https://developer.nvidia.com/tensorrt, 2022.