An Event-driven Inference Approach for Space Computing

Takeharu Eda, Ichibe Naito, Ikuo Yamasaki, Keiichi Tabata, and Xu Shi

Abstract

NTT Software Innovation Center (SIC) has started research and development on a space computing platform on the basis of its experiences in cloud computing and artificial intelligence (AI) inference platforms. This article introduces SIC's views on space-computing-platform requirements, related technologies, and an evaluation of an event-driven AI inference in space computing. As a result of preliminary experimentation, we found that our event-driven AI inference technique may reduce latency and transfer-data size and achieve almost real-time analysis service using observed data in space by further optimizing and improving event-detection models for space.

Keywords: space computing, space datacenter, AI inference, event-driven data processing

1. Toward space computing platform

With the development of remote sensing technologies from Earth observation satellites, it has become possible to observe the ground from space and conduct advanced analysis using artificial intelligence (AI). This technology is being used in a variety of use cases using observation data that cannot be acquired on the ground. However, due to the limited number of available satellites and huge volume of observation data, it often takes several days before actual analysis becomes possible, making it difficult to provide real-time services. To tackle these challenges, NTT aims to construct a *space datacenter*, which is a space computing platform where we can analyze observation data in real time in space and send only important information to the ground using a high-speed transmission network service, reducing unnecessary data transfers and latency.

NTT Software Innovation Center (SIC) has started research and development on a space computing platform on the basis of its experiences in cloud computing [1] and AI inference platforms [2]. This article introduces SIC's perspective on space-computing-platform requirements and related technologies.

2. Technical challenges in space computing

Space computing literally means computing in space, but there are many differences in assumptions compared with computing on the ground. Since the amount of available power in space is limited by the capacity to generate electricity with solar panels, computers in space are required to be power-efficient. Once a computer is launched into space, it is not easy to repair it, so high reliability and stability are required. We also need countermeasures against space-specific difficulties such as radiation, inability to use air for cooling due to the vacuum, microgravity, and temperature. Due to the development and verification time required against these challenges, computers currently in space are several generations older and their performance is significantly lower than that of computers on the ground. These hardware limitations have made it difficult to develop software, forcing us to program in a low-level environment as in embedded systems. However, there have been efforts to tackle these technical challenges for space computing.

Server manufacturers and chip venders have started experiments in space with their commercial off-the-shelf products. For example, Hewlett Packard
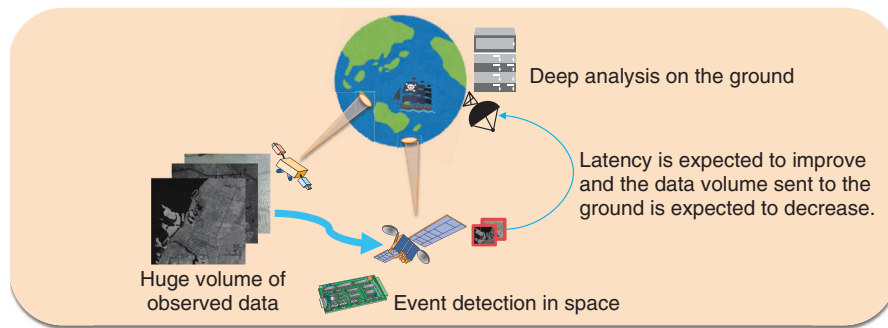
Fig. 1.   Event-driven data processing in space.

Enterprise's Spaceborne Computer, which is based on the ProLiant series of computers on the ground, is installed on the International Space Station (ISS) and connected to the cloud environment on the ground to conduct various experiments such as data analysis with AI [3]. Intel is experimenting with commercial AI chips on a small satellite project called PhiSat-1 [4] for data processing in space. Such efforts aimed at providing a computing environment in space are expected to result in more advanced data processing in space as the current hardware becomes more powerful.

Along with hardware development, there is a trend to adopt modern and efficient software stacks and development/testing techniques in space, and Japan Aerospace Exploration Agency (JAXA) has launched the Satellite DX Research Group to promote satellite digital transformation (DX), which improves flexibility and reduces development cost and time with the "softwarization of satellites" [5].

With the advances in hardware and software development environments, there are growing expectations for new services that extract useful information from observed data in space and deliver it to users in real time through advanced data analysis with AI. However, even with the advances in space computing technologies, there will continue to be many cases in which complete data analysis is impossible in space. Since the data taken by remote sensors, such as a synthetic aperture radar and optical cameras, have high resolution and large data volume (several gigabits (GB) per image, several hundred GB per day), it takes time to transmit the data to the ground even when using fast optical communication. The cost of data analysis on the ground for processing and saving is also expected to increase.

3.   Event-driven data processing in space

SIC has been developing event-driven inference technology for AI services targeting on the ground [2]. In an event-driven inference, a lightweight inference, such as event detection, is carried out on the device side, and only data requiring high-precision processing are sent to the server, thus reducing the amount of transfer. This is expected to effectively reduce the amount of computation, processing, and power consumption for AI analysis. We verified whether such an event-driven inference is also effective in space computing (**Fig. 1**).

For this verification, we assumed a use case of suspicious ship detection, and conducted the verification on the basis of the assumption that a large amount of GB-class image data were captured using an optical remote sensor in space. Relatively lightweight event detection can be carried out in the low-spec computing environment of space, and only data that need to be processed on the ground are transmitted, thus reducing the amount of data transfer. On the ground, we conduct highly accurate AI processing only on the data sent to us to reduce the amount of computation.

Figure 2 shows the implemented data-processing pipeline and elapsed time for each processing. We implemented a case in which a large image is divided into a set of small patches followed by event-detection processing for all patches in space, while on the ground, deep analysis with accurate AI models is conducted on the transmitted data. Event detection, which requires processing tens of thousands of images after division, was carried out using lightweight algorithms, such as histogram-based filtering algorithms with predefined thresholds for detecting events, because of the limited computing power in space.

Fig. 2. Data processing pipeline and elapsed time for each processing.
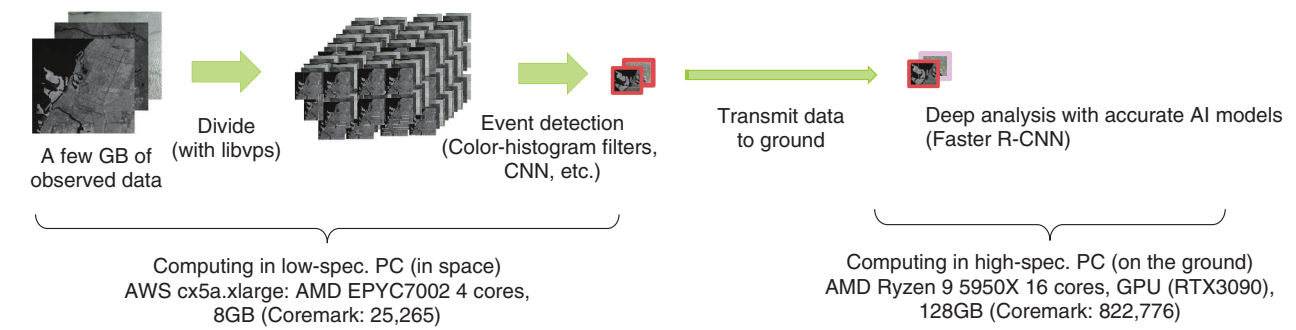
	Divide	Event detection		Deep analysis
Processing time [min.]	4.5	3.7		10.5–11.3

GPU: graphics processing unit
PC: personal computer

Table 1. Recall rates of event detection and reduction rates of data transfer using lightweight filtering algorithms with two thresholds.

Algorithm	Recall rate of event detection	Reduction rate of transfer data	Elapsed time
Color-histogram filtering 1	75%	7.1%	2.17 min. for 10,000 images
Color-histogram filtering 2	70%	14.3%	

4. Results and discussion

In the testing environment for space computing, we confirmed that the entire processing of large observed data can be completed in about 10 minutes with proper resource utilization. Since processing on the ground can be easily parallelized with computing clusters, we can say that real-time AI inference is feasible. However, accuracy and transfer-data reduction were not satisfactory in this experiment. Generally speaking, there is a trade-off between accuracy (in this case, recall) and transfer-data reduction, but even with a 70% recall rate, only 14.3% of the transfer volume could be reduced (**Table 1**). This is due to the low accuracy of the lightweight filtering algorithms used for event detection in this setting.

We then calculated the recall rates of event detection and reduction rates of data transfer on the assumption that highly accurate AI models for deep analysis on the ground are currently in space (**Table 2**).

If there are accurate AI models (e.g., Mask R-CNN) in space, we can reduce 55.0 to 65.7% of transfer data with little accuracy degradation, but the central processing unit in our testing environment makes this almost impossible, requiring us to further improve the speed and optimization for specialized accelerators in the space environment.

5. Future direction of space computing research in SIC

On the basis of the results of the study we presented, we believe that the use of advanced AI accelerators and a system-on-a-chip with various accelerators on a single chip is essential to achieve highly efficient event-driven processing in space computing. SIC has been conducting research and development on benchmarking of AI accelerators for video analysis and optimization of AI models using an open source AI compiler (Apache TVM), which has a mechanism suitable for hetero-architecture. We will use this

Table 2. Recall rates of event detection and reduction rates of data transfer using highly accurate AI models with two thresholds.

AI model	Recall rate of event detection	Reduction rate of transfer data	Elapsed time
Mask R-CNN 1	88.7%	55.0%	22 hours and 16 min.
Mask R-CNN 2	72.2%	65.7%	

knowledge and experience in the research and development of a space computing platform and develop more advanced AI inference techniques on the basis of event-driven data processing for space.

References

[1] "Feature Articles: R&D Efforts in Cloud Computing Platform Technologies through Open Innovation," NTT Technical Review, Vol. 13, No. 2, 2015.
https://www.ntt-review.jp/archive/2015/201502.html
[2] T. Eda, R. Kurebayashi, X. Shi, S. Enomoto, K. Iida, and D. Hamuro, "An Efficient Event-driven Inference Approach to Support AI Applications in IOWN Era," NTT Technical Review, Vol. 19, No. 2, 2021.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202102fa4.html
[3] Spaceborne Computer-2 High Performance Commercial Off-The-Shelf (COTS) Computer System on the ISS, https://www.nasa.gov/mission_pages/station/research/experiments/explorer/Investigation.html?#id=8221
[4] Φ-sat, https://www.esa.int/Applications/Observing_the_Earth/Ph-sat
[5] Research Group for Satellite DX (Digital Transformation), JAXA (in Japanese), https://www.kenkai.jaxa.jp/research/sasshin/sasshin.html

Trademark notes

Takeharu Eda

Senior Research Engineer, AI Application Platform Project, NTT Software Innovation Center.

He received a B.S. in mathematics from Kyoto University in 2001 and M.S. in engineering from Nara Institute of Science and Technology in 2003. He joined NTT in 2003, and his research interests include a wide range of topics in machine learning and systems. He is a member of the Information Processing Society of Japan (IPSJ) and the Association for Computing Machinery (ACM).

Keiichi Tabata

Researcher, NTT Software Innovation Center.

He received a B.E. and M.E. in computer science and engineering from Waseda University, Tokyo, in 2010 and 2012. He joined NTT in 2012, and his research interests include computer architecture, compiler optimization, software engineering, and deep learning. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), IEEE Computer Society, IEEE Communications Society, ACM, the Institute of Electronics, Information and Communication Engineers (IEICE) and IPSJ.

Ichibe Naito

Manager, Engineering, Space DC Business, Space Compass Corporation.

He joined NTT in 2006, and his research interests include distributed systems. In 2010, he moved to NTT Communications where he designed the database of the operation support system and developed an IaaS (Infrastructure as a Service) service. In 2013, he joined NTT Software Innovation Center and developed an operation support system for distributed object storage. In 2022, he joined Space Compass, which takes on the challenge of building new infrastructures in space. He is currently engaged in the development of edge computing in space.

Xu Shi

Engineer, AI Application Platform Project, NTT Software Innovation Center.

She joined NTT in 2014. Her current research interests include deep learning and computer vision.
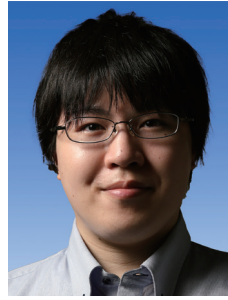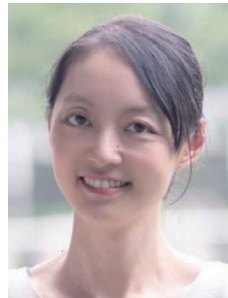
Ikuo Yamasaki

Senior Research Engineer, Supervisor, AI Application Platform Project, NTT Software Innovation Center.

He received a B.S. and M.S. in electronic engineering from the University of Tokyo in 1996 and 1998. He joined NTT in 1998 and was involved in research and development (R&D) activities regarding OSGi. In 2006, he was a visiting researcher at IBM Ottawa Software Laboratories, IBM Canada. He is currently leading and managing R&D activities related to the AI application platform at NTT laboratories.