INAF
ISTITUTO NAZIONALE
DI ASTROFISICA

OSSERVATORIO DI ASTROFISICA
E SCIENZA DELLO SPAZIO DI BOLOGNA

OAS
BOLOGNA

Onboard AI: Enhancing High-Energy Space Missions with Machine Learning

N. Parmiggiani, A. Bulgarelli, L. Castaldini, D. Ammirabile, R. Falco, A. Di Piano
INAF/OAS Bologna

AI In Astronomy - Workshop
21 May 2025

Outline

- Introduction of the Onboard AI context
- Examples of DL models that we developed for AGILE
- Prototype setup with scintilators detectors and Edge Computers
- Optimization techniques for DL models

Context

RTA group @ INAF OAS/bologna

- X- and gamma-ray astrophysics expertise in:
  - **Development of real-time analysis pipelines** for identification of gamma-ray transients in space missions (AGILE, COSI) and ground arrays of telescopes (CTA, ASTRI Mini-Array)
  - **Monte Carlo simulations** of scientific payloads: ASTROGAM, AGILE, COSI, Athena
  - **Development of machine learning algorithms** for gamma-ray transient detection and localisation (AGILE, CTA, COSI)
  - **EdgeAI** studies for future space mission

Future of scientific space missions

- New sensitive detectors to collect **ever-increasing amounts of data**.
- Goal of scientific space missions:
  - Observe in greater detail and depth **known sources**
  - Detect **transient phenomena in the Universe**
    - we don't know **when and where** they will happen
  - Observe **the unexpected**
    - we don't know **what, where and when** we will observe

How Onboard AI can contribute

- **Flexibility:** update the onboard analysis software to adapt it with the new scientific requirements that can occur during the mission lifetime

- **Spacecraft autonomy**: detect and localize transient sources. Repoint the telescope and optimize the synergies between onboard instruments.

- **Data transmission optimization:** reduce the data volume that is transmitted to the ground.

- **Health monitoring:** perform health monitoring analysis directly onboard without the need to send to the Earth monitoring data.

Improve the flexibility with AI

- The **Scientific Requirements of a space mission evolves** during time, considering the long duration of these projects.

- We had direct experience with this being part of the AGILE space mission team. After the detection of Gravitational Waves GW (2015) the AGILE agile team **updated the onboard configuration of AGILE** to improve the detection capabilities of transient events.

- The standard procedure to update onboard configuration of a satellite is very complex and limited by the original design of the mission. The usage of onboard AI to perform data analysis allows for a **faster and easier update of onboard software**, for example by **uploading new models** with new features.

Improve the Spacecraft Autonomy

- In the multi-messenger and multi-wavelength era the **fast response to transient events** (such as Gamma-ray Bursts) is crucial to study these phenomena.

- During last years the space missions improved the response time to transient events from hours to seconds. To achieve this results it is necessary to have **automated onboard analysis systems** avoiding the needs for ground communications.

- Using **onboard ML and DL analysis techniques can improve the reaction time** to transient events. Indeed a fast detection and localization is mandatory to repoint the instruments or to send a science alert to the scientific community.

Impact on the Ground Segment

- **Onboard AI can pre-process, filter and compress data** onboard to reduce the volume of data transmitted to the ground. Onboard analysis can classify data and discard low quality data. This mean a **cost reduction** for the ground segment.

- Onboard AI algorithm can be used to **optimize the observation schedule**, reducing human intervention from the ground. It is also possible to optimize the data management considering the communication time windows.

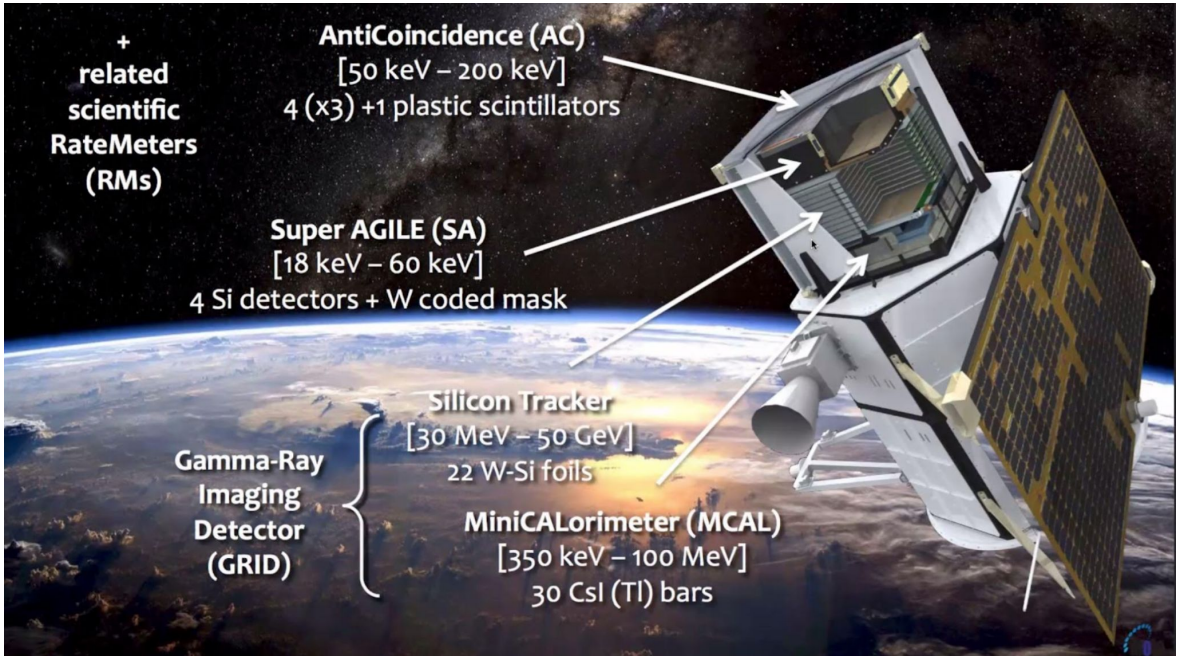
- The **S/C health can be monitored by onboard AI algorithm** to detect anomalies in real-time without the need to send data to the Earth.

- **The Ground Segment has to integrate the AI model lifecycle**. The model can be trained on the ground or directly onboard if the computing resources are enough.

DL models to analyze gamma-ray data

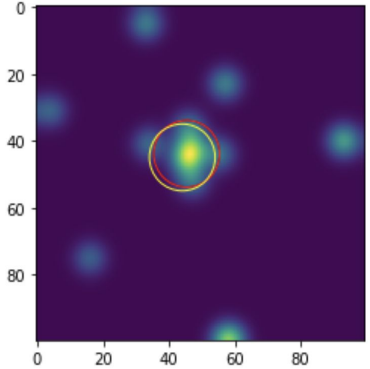
AGILE satellite

- **AGILE** is an ASI space mission launched in 2007, designed to study X-ray and gamma-ray astronomy. AGILE terminated the in-orbit operations on February 2024, after almost 17 years of successful scientific observations.



Detection and localization of GRBs in Sky Maps

- We developed two **Convolutional Neural Networks** (CNN) to **detect and localize** GRBs from the AGILE/GRID **counts maps**.
- We simulated three **datasets with 40 000 maps** for the training, testing, and validation phases. The CNN is trained with a **supervised learning technique**, so the datasets are labeled.
- We evaluated the CNN using the catalogs of other space missions detecting **21 GRBs with a sigma > 3** while the standard **Aperture Photometry** analysis detects **only two GRBs.**
- The localization is executed with a CNN to perform a **regression** on the coordinates of the GRBs with a **mean error of 0.7 degrees**.

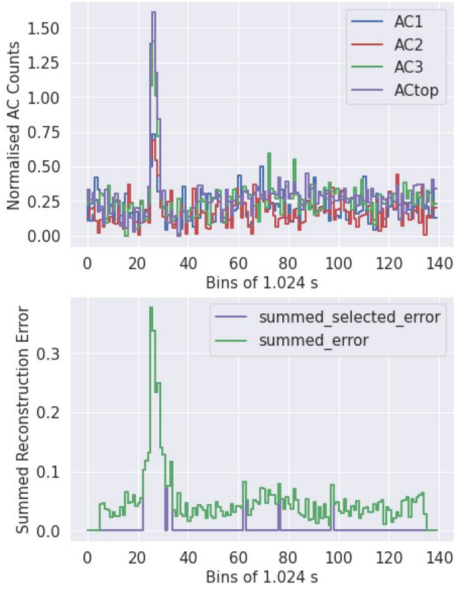


Parmiggiani N. et al. A Deep Learning Method for AGILE/GRID Gamma-ray Bursts detection, Astrophysical Journal, Volume 914, Issue 1, id.67, 12 pp (2021)

Parmiggiani N. et al. Preliminary Results of a New Deep Learning Method to Detect and Localize GRBs in the AGILE/GRID Sky Maps. proceedings of the ADASS XXXII (2022) conference arXiv

Anomaly Detection in time series

- We implemented a **Convolutional Neural Network autoencoder** to **detect GRBs** in the ratemeters of the AGILE Anticoincidence System (ACS).
- The autoencoder aims to reconstruct the input data, **minimizing the reconstruction error,** and can be used for **anomaly detection.**
- We trained the model using an **unsupervised technique** using 5000 background-only time series.
- We evaluated the trained model using GRBs detected by other space missions, and the model **detected 72 GRBs**, **15 of which** were detected for the first time in the AGILE data.



Parmiggiani N., Bulgarelli A., Ursi A. et al. “A Deep-learning Anomaly-detection Method to Identify Gamma-Ray Bursts in the Ratemeters of the AGILE Anticoincidence System”, Astrophysical Journal, Volume 945, (2023)

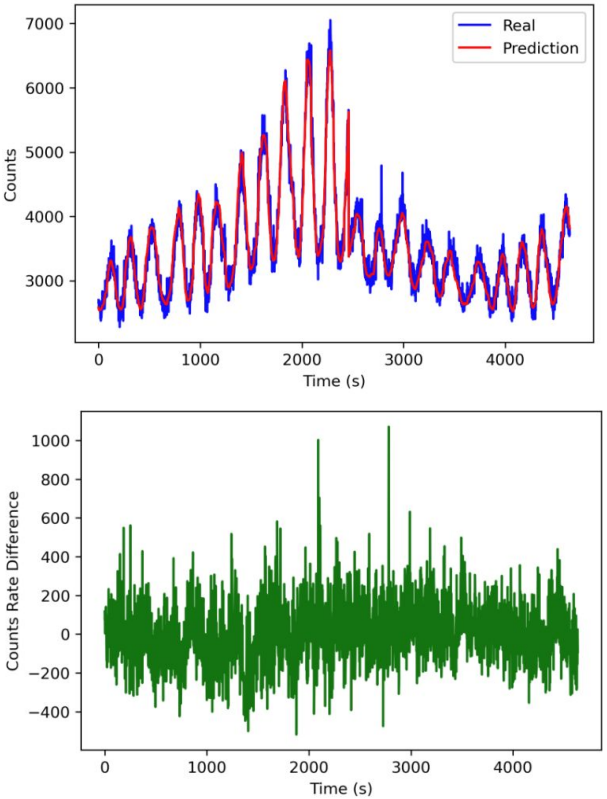
Deep Learning to predict the ACS background

- **Goal**: predict the background count rates of the AGILE ACS system using the satellite **orbital and attitude parameters.**

- We calculated the **difference** between the **real** and **predicted** counts of the test dataset to check the accuracy of the model. The model has a mean prediction error of 3.8%.

- We can use the predicted counts of the background **to detect GRBs** where the differences with the acquired counts are higher than a predefined threshold.

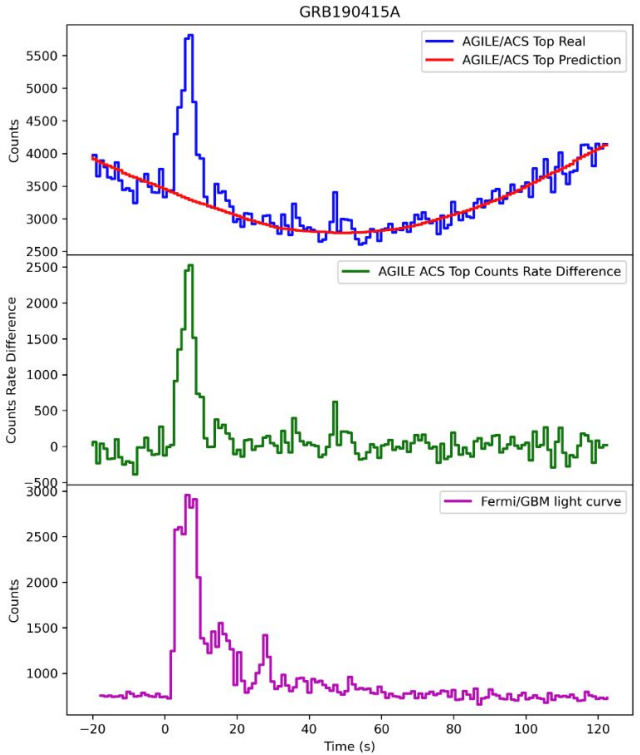
- We can apply this detection method to raw data **without applying the detrending algorithm** that can introduce artificial anomalies.



Similar approach used by R. Crupi et al. *Searching for long faint astronomical high energy transients: a data driven approach.* EA 2023

Detect GRBs using the predicted values

- We tested this detection method using the **GRB web catalog** and extracting light curves from the ACS archive (2019-2022).
- The method **detects 39 GRBs** with sigma > 3. **Four GRBs are new detections** that were not detected in previous analyses.
- We also compared the results obtained with the **light curve of the Fermi/GBM** detector because they have a similar energy range. ACS (50-200 keV) and Fermi/GBM (50-300 keV)
- We are investigating other possible applications of this kind of Deep Learning model to predict the background level of the AGILE detectors.



Parmiggiani N., et al. “A New Deep Learning Model to Detect Gamma-Ray Bursts in the AGILE Anticoincidence System”, Astrophysical Journal, Volume 973, (2024)

Edge Computing and Onboard AI

Edge Computers and AI

- Edge computers are **low power consumption devices** that process data near the acquisition hardware instead of collect all data in a centralized data center.
- Thanks to their low power consumption and dimension they can be **deployed onboard satellites** and cubesats.
- **Goal: optimize ML/DL models to be executed on Edge Computers**

Redpitaya

- Dual-core ARM Cortex-A9, 372 MFLOP/core
- FPGA: Xilinx Zynq 7020 (can be used for AI)
- 2 ADC channel (125 Mbps)
- RAM LPDDR4 512 MB
- 5-10 W

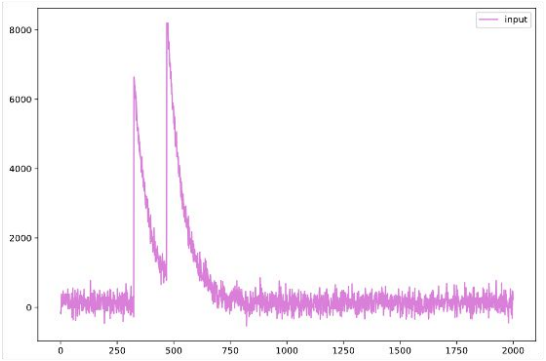
Jetson Orin Nano

- CPU Arm® Cortex®-A78AE v8.2 6-core 64-bit
- 1,5 MB L2 + 4 MB L3
- GPU NVIDIA Ampere a 1024, 32 Tensor Core
- RAM LPDDR5 8 GB
- 7-15 W

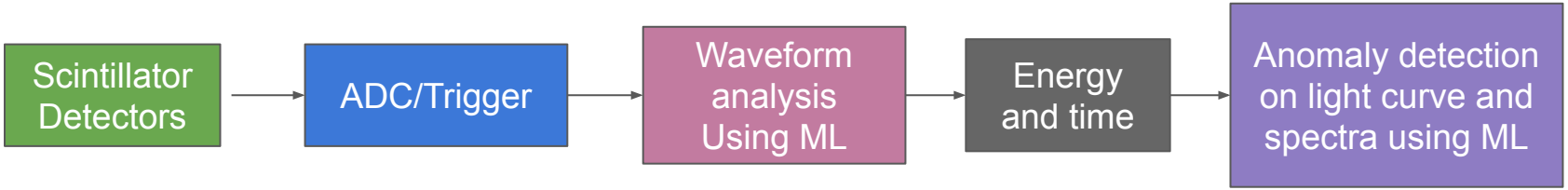
Jetson AGX Orin

- CPU Arm® Cortex®-A78AE v8.2 12-core 64-bit 3 MB L2 + 6 MB L3
- GPU NVIDIA Ampere, 2048 core , 64 Tensor Core
- RAM LPDDR5 64 GB
- 15-60 W

Experimental Setup to teset Edge AI



2) The waveform are analyzed with ML to obtain the energy and time of the photons. Then the photon list is analyzed with anomaly detection ML models.

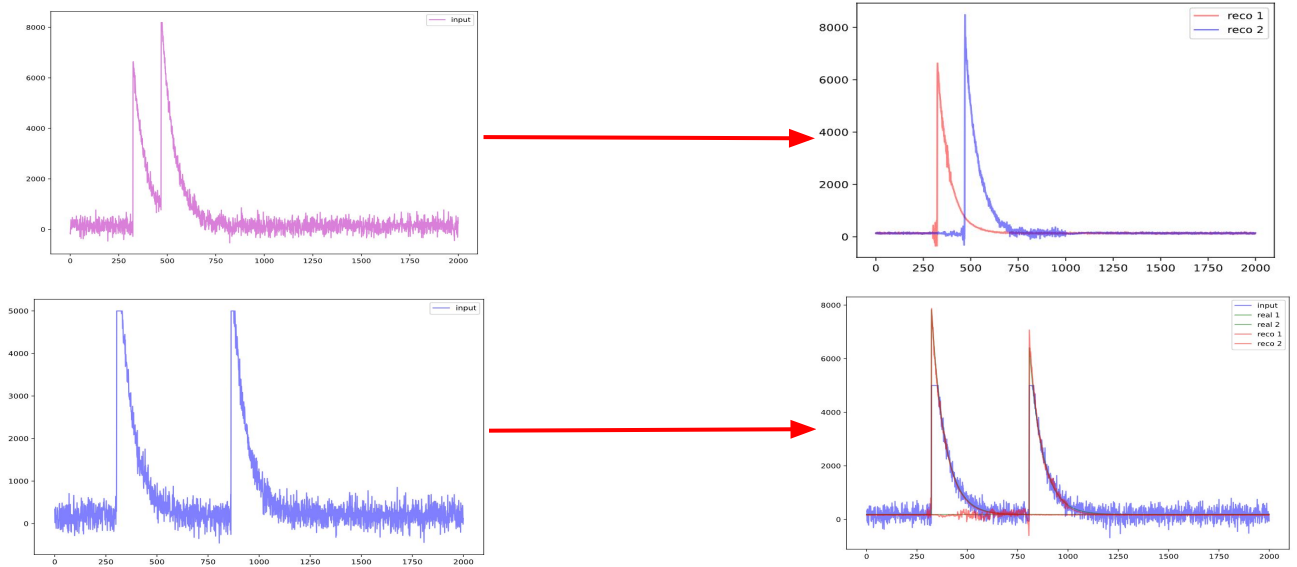


1) The data is acquired usinga Redpitaya ADC, then a trigger algorithm deployed on the FPGA detect the waveform with a signal that must be analyzed.

Scintillator Detectors' Waveform Analysis

- We developed a **Convolutional Neural Network** with three 1D layers and a output fully connected layer to **separate the piled up waveforms** and **reconstruct saturated waveform**

- This model can be used onboard satellites to improve scintillator detector efficiency



Deep Learning Models optimization

- We are evaluating optimization techniques to **reduce the size of trained DL models and their inference time**. These optimizations are critical to deploy the models on low-power Edge Computing devices onboard satellites and CubeSats.

- We can use **Pruning** and **Quantization** techniques to improve our models:

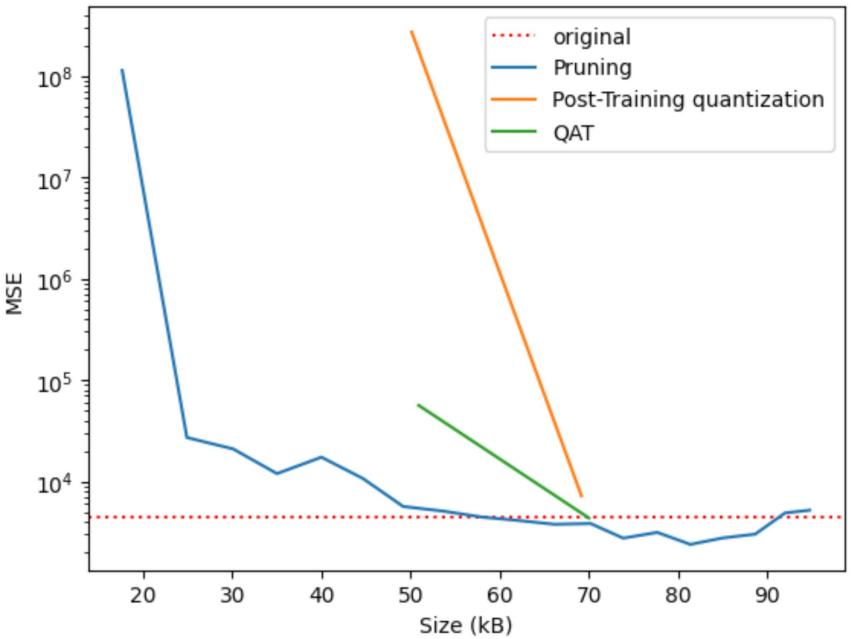
    - **Pruning:** continues the model training to remove the unimportant weights from the model gradually increasing the sparsity (percentage of weights set to zero) maintaining accuracy.

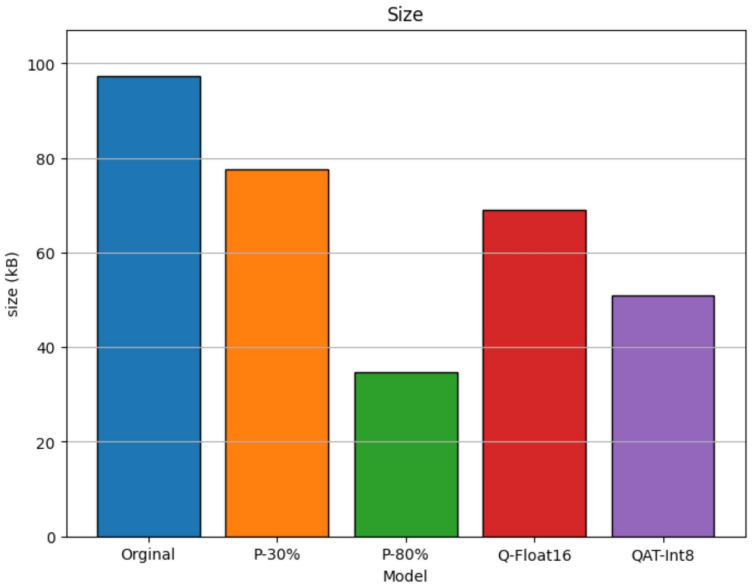
    - **Quantization**: reduces the memory footprint and computational cost of DL models by representing weights and activations with lower precision (e.g. 8-bit or 16-bit float)

        - **Post-Training Quantization**: converts a pre-trained model to lower precision after training
        - **Quantization-Aware Training**: apply quantization during training to minimize accuracy degradation

Preliminary Results

- We have to find the **best trade-off** between size reduction and performance degradation. The **pruning** technique seems to be the best one.





Work in progress

- Use **Tensorflow Lite (LiteRT)** to perform inference with trained model on Edge Computing Devices.
- Improve the inference time using C/C++ and optimization libraries such as **XNNPACK**
- Evaluation of **Quantization and Pruning** improvements and execution of performance tests on real Edge Devices such as the NVIDIA Jetson and other boards.
- Test other optimization techniques such as **Model Distillation**
    - **Teacher model**: complex model with high accuracy
    - **Student model**: learns not just from the labels but from the output of the teacher model

Conclusions

Conclusions

- The next generation of scientific space missions will require **a different data analysis approach**. **We propose the usage of Onboad AI exploiting Edge Computing devices capabilities.**

- Onboard AI can improve the data analysis **flexibility**, the **spacecraft autonomy**, the **data volume reduction**, and the **real-time follow-up of transient events** such as gamma-ray bursts.

- We developed a prototype setup to test the proposed technologies evaluating optimization techniques to **run ML models onboard satellites and CubeSats**.

- We already developed **Deep Learning models to detect and localize GRBs in the sky maps and time series** that can be optimized and deployed on satellites.