# 7th International Workshop on
# On-Board Payload Data Compression

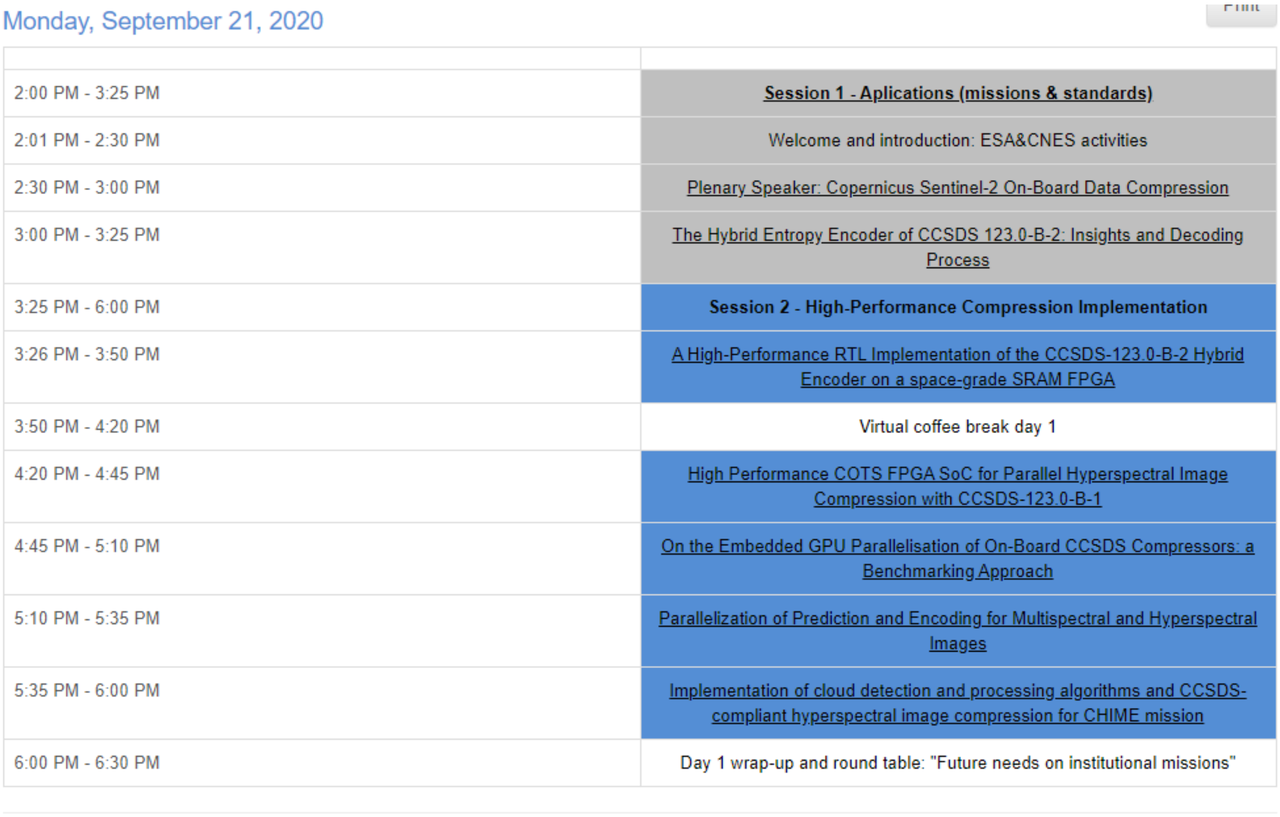
## OBPDC 2020

## 21 - 23 September 2020
## Online Event

# Abstract Booklet

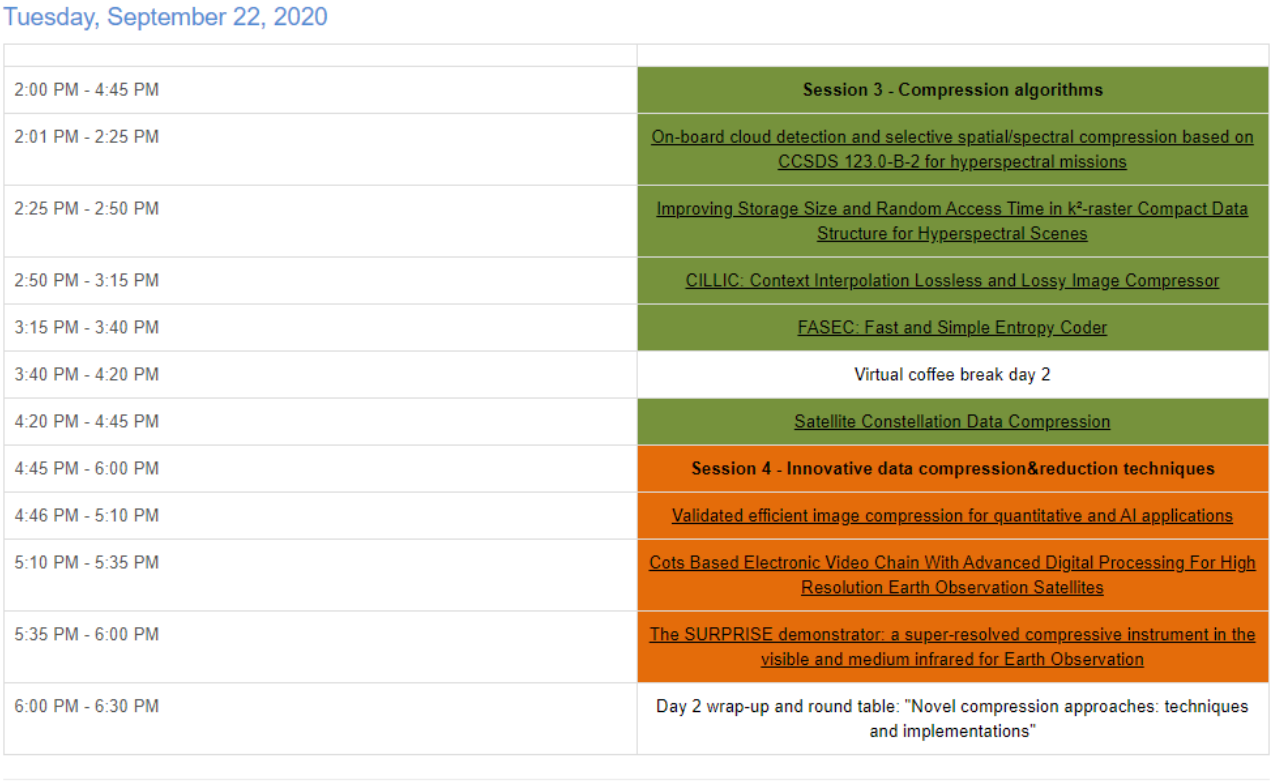
# Contents

# Final Programme



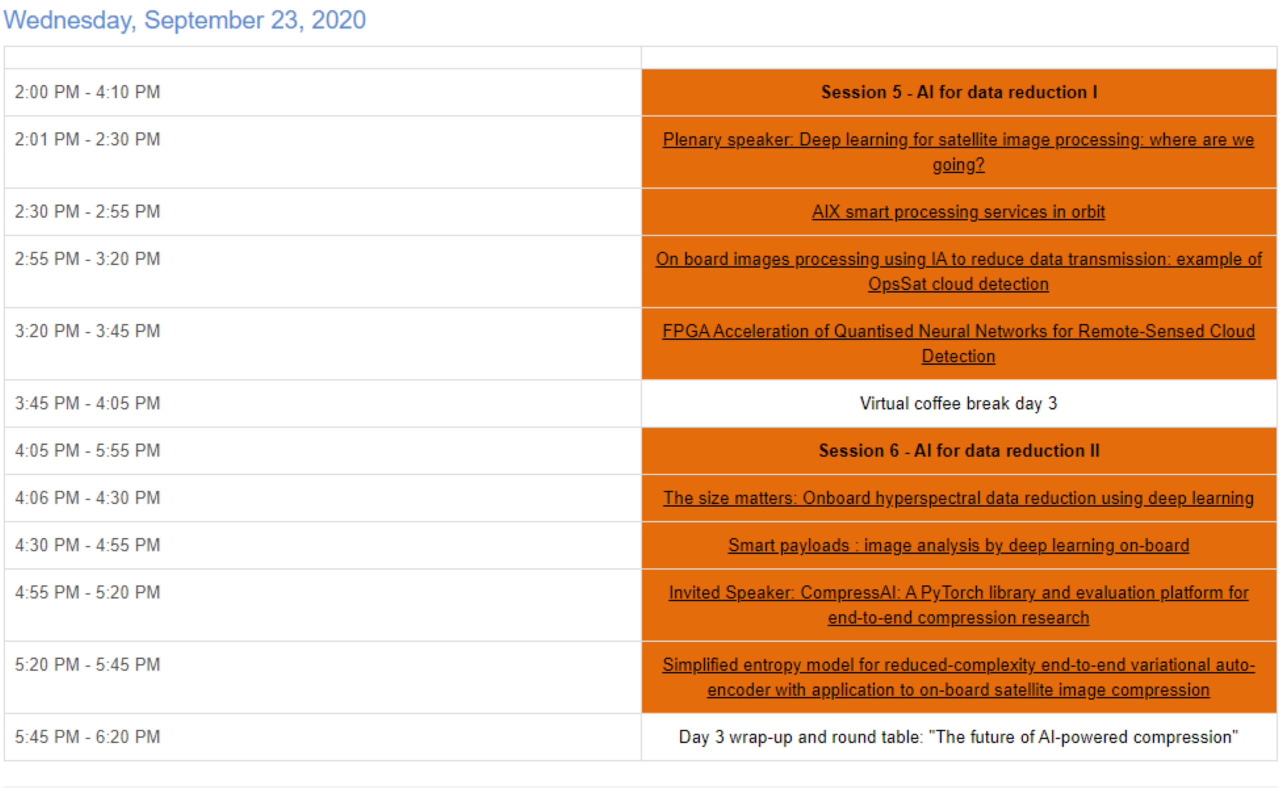
## Monday, September 21, 2020

| | |
|---|---|
| 2:00 PM - 3:25 PM | **Session 1 - Aplications (missions & standards)** |
| 2:01 PM - 2:30 PM | Welcome and introduction: ESA&CNES activities |
| 2:30 PM - 3:00 PM | Plenary Speaker: Copernicus Sentinel-2 On-Board Data Compression |
| 3:00 PM - 3:25 PM | The Hybrid Entropy Encoder of CCSDS 123.0-B-2: Insights and Decoding Process |
| 3:25 PM - 6:00 PM | **Session 2 - High-Performance Compression Implementation** |
| 3:26 PM - 3:50 PM | A High-Performance RTL Implementation of the CCSDS-123.0-B-2 Hybrid Encoder on a space-grade SRAM FPGA |
| 3:50 PM - 4:20 PM | Virtual coffee break day 1 |
| 4:20 PM - 4:45 PM | High Performance COTS FPGA SoC for Parallel Hyperspectral Image Compression with CCSDS-123.0-B-1 |
| 4:45 PM - 5:10 PM | On the Embedded GPU Parallelisation of On-Board CCSDS Compressors: a Benchmarking Approach |
| 5:10 PM - 5:35 PM | Parallelization of Prediction and Encoding for Multispectral and Hyperspectral Images |
| 5:35 PM - 6:00 PM | Implementation of cloud detection and processing algorithms and CCSDS-compliant hyperspectral image compression for CHIME mission |
| 6:00 PM - 6:30 PM | Day 1 wrap-up and round table: "Future needs on institutional missions" |



## Tuesday, September 22, 2020

| | |
|---|---|
| 2:00 PM - 4:45 PM | **Session 3 - Compression algorithms** |
| 2:01 PM - 2:25 PM | On-board cloud detection and selective spatial/spectral compression based on CCSDS 123.0-B-2 for hyperspectral missions |
| 2:25 PM - 2:50 PM | Improving Storage Size and Random Access Time in k²-raster Compact Data Structure for Hyperspectral Scenes |
| 2:50 PM - 3:15 PM | CILLIC: Context Interpolation Lossless and Lossy Image Compressor |
| 3:15 PM - 3:40 PM | FASEC: Fast and Simple Entropy Coder |
| 3:40 PM - 4:20 PM | Virtual coffee break day 2 |
| 4:20 PM - 4:45 PM | Satellite Constellation Data Compression |
| 4:45 PM - 6:00 PM | **Session 4 - Innovative data compression&reduction techniques** |
| 4:46 PM - 5:10 PM | Validated efficient image compression for quantitative and AI applications |
| 5:10 PM - 5:35 PM | Cots Based Electronic Video Chain With Advanced Digital Processing For High Resolution Earth Observation Satellites |
| 5:35 PM - 6:00 PM | The SURPRISE demonstrator: a super-resolved compressive instrument in the visible and medium infrared for Earth Observation |
| 6:00 PM - 6:30 PM | Day 2 wrap-up and round table: "Novel compression approaches: techniques and implementations" |



## Wednesday, September 23, 2020

| | |
|---|---|
| 2:00 PM - 4:10 PM | **Session 5 - AI for data reduction I** |
| 2:01 PM - 2:30 PM | Plenary speaker: Deep learning for satellite image processing: where are we going? |
| 2:30 PM - 2:55 PM | AIX smart processing services in orbit |
| 2:55 PM - 3:20 PM | On board images processing using IA to reduce data transmission: example of OpsSat cloud detection |
| 3:20 PM - 3:45 PM | FPGA Acceleration of Quantised Neural Networks for Remote-Sensed Cloud Detection |
| 3:45 PM - 4:05 PM | Virtual coffee break day 3 |
| 4:05 PM - 5:55 PM | **Session 6 - AI for data reduction II** |
| 4:06 PM - 4:30 PM | The size matters: Onboard hyperspectral data reduction using deep learning |
| 4:30 PM - 4:55 PM | Smart payloads : image analysis by deep learning on-board |
| 4:55 PM - 5:20 PM | Invited Speaker: CompressAI: A PyTorch library and evaluation platform for end-to-end compression research |
| 5:20 PM - 5:45 PM | Simplified entropy model for reduced-complexity end-to-end variational auto-encoder with application to on-board satellite image compression |
| 5:45 PM - 6:20 PM | Day 3 wrap-up and round table: "The future of AI-powered compression" |

# Abstracts – Oral Presentations

## Session 1 - Applications (missions & standards)

### Welcome and introduction: ESA & CNES activities

**Speaker TBC**

No abstract available.

*************************************************************

### Plenary Speaker: Copernicus Sentinel-2 On-Board Data Compression

Dr. Ferran Gascon[1]

[1]*European Space Agency (ESA), Frascati, Italy*

This talk will include a description of the Copernicus Sentinel-2 on-board data compression system and the operational experience of using it.

*************************************************************

### The Hybrid Entropy Encoder of CCSDS 123.0-B-2: Insights and Decoding Process

Ian Blanes[1], Aaron Kiely[2], Lucana Santos[3], Miguel Hernández-Cabronero[1], Joan Serra-Sagristà[1]

[1]*Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain,* [2]*NASA Jet Propulsion Laboratory (JPL), Pasadena, United Stated of America,* [3]*ESA ESTEC, Noordwijk, Netherlands*

In February 2019, the Consultative Committee for Space Data Systems (CCSDS) published the 123.0-B-2 standard titled “Low-Complexity Lossless and Near-Lossless Multispectral and Hyperspectral Image Compression”. It included a new state-of-the-art entropy encoder: the hybrid entropy encoder. This encoder, originally proposed by NASA’s Jet Propulsion Laboratory (JPL), was designed to efficiently code the low entropy data produced by the in-loop quantizer included in the standard. This document describes some of the insights that went into the design of that entropy encoder.

It is well known that in order to have an efficient entropy encoder at low entropies, the information of multiple symbols needs to necessarily be combined into a single codeword. Otherwise, compressed data rates cannot be lower than 1 bit per sample. The hybrid encoder in 123.0-B-2 uses a set of variable-to-variable (V2V) codes. These complement the Golomb-power-of-two (GPO2) codes, which are more efficient for for higher entropy data, and were already used in previous CCSDS compression standards.V2V codes are used for low entropies, while GPO2 codes can still be employed for higher entropies. This, however, presents some technical issues.

For example, the use of V2V codes in combination with an adaptive probability model, poses an interesting challenge: it is not possible for the probability model to work directly with the V2V codes. This is because a regular decoder needs to update the probability model using symbols that the V2V codes in the encoder are still packing together, creating a Catch-22 situation further described in the manuscript. This problem was finally solved by reconfiguring the compressed data so that it can be decoded in the reverse order. This design decision not only addresses the issue, but that at the same time moves the hardware cost of addressing this situation from the space-borne encoder to the ground decoder.

Another design decision described in this document is the use of a "breadcrumbing" strategy. The encoder leaves small pieces of information at critical points in the compressed file, allowing the decoder to retrace the encoder steps but in a reverse order. In 123.0-B-2, the final state of the encoder is transmitted at the

end of a compressed image so that the decoding process can begin exactly where coding stopped. In addition, the adaptive probability model is kept in sync while decoding thanks to small pieces of information embedded in the compressed file that disambiguate rounding operations.

A few insights into the design of the V2V codes themselves are included as well. For instance, a data-driven model was involved in the optimization of the V2V codes that were finally standardized. This is in contrast to probability models based on geometric distributions, which are a sufficiently well fit for GPO2 codes. For V2V codes, the use of a data-driven model yielded gains worthwhile the effort.

In addition to providing some of the design insights that went into the design of the hybrid encoder in CCSDS 123.0-B-2, this paper aims to clearly describe the decoding process. This complements the CCSDS 123.0-B-2 standard, which only describes the operations from the point of view of a compliant compressor. Due to the notable differences in the coding and decoding processes of the hybrid entropy encoder, this manuscript provides as well a straightforward description of the steps involved in the decoding process.

*********************************************************

## Session 2 – High-Performance Compression Implementation

### A High-Performance RTL Implementation of the CCSDS-123.0-B-2 Hybrid Encoder on a space-grade SRAM FPGA

Mr. Panagiotis Chatziantoniou[1], Mr. Antonis Tsigkanos[1], Prof. Nektarios Kranitis[1,2]

*[1]Digital Systems and CCILomputer Architecture Laboratory (DSCAL), National and Kapodistrian University of Athens, Athens, Greece, [2]University of the Peloponnese, Sparta, Greece*

Nowadays, hyperspectral imaging is recognized as a cornerstone remote sensing technology. The latest high-resolution and high-speed space-borne imagers have brought an explosive growth in data volume and instrument data rates in the range of several Gbps. This competes with the limited on-board storage resources and downlink bandwidth, making hyperspectral image data compression a mission critical on-board processing task. Due to the high data volume reduction often needed to meet spacecraft downlink bandwidth requirements, lossy compression is becoming increasingly important. In this context, the Multispectral Hyperspectral Data Compression (SLS-MHDC) Working Group of the Consultative Committee for Space Data Systems (CCSDS) standardized the new Issue 2 "Low-Complexity Lossless and Near-Lossless Multispectral and Hyperspectral Image Compression" standard CCSDS-123.0-B-2. This new Issue 2 extends Issue 1, incorporating support for low-complexity near-lossless compression, while retaining compatible lossless compression capabilities, where “near-lossless” refers to the ability to perform compression in a way that limits the maximum error in the reconstructed image to a user-specified bound.

A key feature of CCSDS-123.0-B-2 is the new Hybrid Encoder option. At high bit-rates, the Hybrid Encoder encodes most samples using a family of codes that are equivalent to those used by the Sample-Adaptive Encoder of Issue 1, and thus has nearly identical performance. However, at low bit rates it has substantially better performance than the Issue 1 entropy encoders.  For example, the Sample-Adaptive Encoder of Issue 1 cannot reach bit-rates lower than 1 bit-per-sample due to design constraints, while the Block-Adaptive Encoder may, but at a non-negligible bit-rate overhead.

The Hybrid Encoder option specified in CCSDS-123.0-B-2 is a modified version of the one originally used by the NASA FLEX entropy coder so that decoding proceeds in reverse order. This permits a more memory-efficient implementation than FLEX’s original coder, which was based on an interleaved entropy coding approach. The Hybrid Encoder includes codes equivalent to the length-limited GPO2 codes used by the Sample-Adaptive Encoder but it is augmented with an additional 16 variable-to-variable length “low-entropy" codes to provide better compression of low-entropy data. Such low-entropy data become more prevalent as increasing predictor quantization step sizes are used. The Hybrid Encoder adaptively switches between high and low entropy encoding methods on a sample-by-sample basis, using code selection statistics similar to those used by the Sample-Adaptive coder. A single output codeword from a low-entropy code may encode multiple samples, which allows obtaining lower compressed data rates than can be produced by the Sample-Adaptive Entropy coder.

In this contribution, we introduce a high-performance hardware implementation of the CCSDS-123.0-B-2 Hybrid Encoder targeting space-grade SRAM FPGA technology, described in portable VHDL RTL. The proposed Hybrid Encoder RTL architecture comprises 6 components at the top structural level: 1) an Adaptive Code Selection Statistics unit, 2) a High/Low Entropy Decision unit, 3) a High Entropy Encoder unit, 4) a Low Entropy Encoder, 5) a Codeword Combiner and 6) a Variable Length Code Packer unit. The proposed architecture achieves 1 sample per cycle when the high-entropy encoding method is selected and multiple cycles per sample when low-entropy coding is selected, while the deep pipeline enables very high clock frequencies.  Moreover, the proposed architecture exploits the systolic design pattern to provide modularity and latency insensitivity in a elastic pipeline.

The implementation of the proposed architecture on a Xilinx Kintex Ultrascale XQRKU060 space-grade SRAM FPGA achieves state-of-the-art throughput performance of up to 354 MSamples/s (5.66 Gbps @ 16bpp) while occupying approximately 2.85% of device LUTs and 0.1% BRAMs of the FPGA resources. To the best of our knowledge, this is the first implementation of the CCSDS-123.0-B-2 Hybrid Encoder Implementation a space-grade SRAM FPGA.

*************************************************************

## High Performance COTS FPGA SoC for Parallel Hyperspectral Image Compression with CCSDS-123.0-B-1

Mr. Antonis Tsigkanos[1], Dr. Nektarios Kranitis[1,2], Mr. Dimitris Theodoropoulos[1], Dr. Antonios Paschalis[1]
*[1]Digital Systems and Computer Architecture Laboratory (DSCAL), National and Kapodistrian University of Athens, Athens, Greece, [2]University of the Peloponnese, Sparta, Greece*

Nowadays, hyperspectral imaging is recognized as a cornerstone remote sensing technology. Next generation, high-speed airborne and space-borne imagers, have increased resolution, resulting in an explosive growth in data volume and instrument data rate in the range of GPixels/s. This competes with limited on-board resources and bandwidth, making hyperspectral image compression a mission critical on-board processing task. At the same time, the "New Space" trend is emerging, where launch costs decrease, and agile approaches are exploited building smallsats using Commercial-Off-The-Shelf (COTS) parts. In this contribution, we introduce a high performance parallel implementation of the CCSDS-123.0-B-1 hyperspectral compression algorithm targeting SRAM FPGA technology. The architecture exploits image segmentation to provide robustness to data corruption and enables scalable throughput performance by leveraging segment-level parallelism. Furthermore, we exploit the capabilities of a COTS FPGA SoC device to optimize SWaP-C.

The architecture partitions a hyperspectral cube stored in a DRAM frame-buffer into segments, compressing them in parallel using a flexible software scheduler hosted in the SoC CPU and several compressor accelerator cores in the FPGA fabric. The input hyperspecral image is split across the Y-axis to form segments compressed independently with default prediction weights.The System-on-Chip is built around the CCSDS 123.0-B-1E IP core, and places a number of those compressors in parallel, scheduling configuration and data movement such that overall throughput scales linearly with the number of placed compressor cores. The software scheduler orchestrates a set of DMA controllers, to segment and transfer image data to the compressors at speed. The embedded ARM processors are used for control tasks only and can therefore be used to perform other preprocessing tasks such as data reduction or classification.

To evaluate the architecture, we perform on-chip benchmarking, instrumenting the control software to time and verify the compression operation of an AVIRIS sensor scene, under different levels of parallelism and segment height. A 5 core implementation demonstrated on a Zynq-7045 FPGA achieves throughput performance of 1387 Msamples/s (22.2 Gbps @ 16bpp), outperforms previous implementations in equivalent FPGA technology, allowing seamless integration with next-generation hyperspectral sensors. This implementation requires a Xilinx Zynq device and external memory, utilizing 91845 LUTs and 448 Block RAMs for a full SoC with 5 cores on the Zynq-7045 device.

The contribution first provides background on the CCSDS 123.0-B-1 algorithm as well as the trade-offs in pixel order, segmentation and considerations for COTS device usage. Benchmarks for various segmentation heights are presented to provide context on the compression performance overhead of segmentation. Then the proposed SoC architecture is described in detail including the parallelism scheme compared to other possible architectures, hardware and software design, as well as considerations on system integration in a payload data chain, consuming sensor data from device I/O and output compressed data towards a downlink. Finally, experimental results are presented, including FPGA implementation results, scaling overhead, technology limitations on scaling and on-chip benchmarked performance. Throughput performance is compared in detail with existing parallel hyperspectral compressors from the literature.

**********************************************************

## On the Embedded GPU Parallelization of On-Board CCSDS Compressors: a Benchmarking Approach

Mr. Iván Rodriguez[1,2], Mr. Álvaro Jover[1,2], Dr. Leonidas Kosmidis[1,2], Mr. David Steenari[3]
*[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain, [2]Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, [3]European Space Agency (ESA), Noordwijk, The Netherlands*

The on-board processing requirements of future missions are constantly increasing, requiring new hardware solutions able to support this need, while staying within the strict power and thermal limits of space systems. Embedded GPUs present a promising candidate, combining high-performance capabilities will low power consumption, close to the target limits. The GPU4S (GPU for Space) ESA-funded project [1] studies whether on-board processing algorithms are amenable to GPU parallelization as well as whether embedded GPUs can satisfy the performance requirements of future space missions, effectively paving the way for their adoption.

Early project results with commonly used processing algorithms [2] as well as an infrared space observatory case study demonstrator [3] indicate that embedded GPUs can provide significant processing improvements of several orders of magnitude compared to existing space processors such as LEON/SPARC or PowerPC-based processors. Compared to FPGAs, which are commonly used in on-board processing applications, GPUs offer the advantage to reconfigure the on-board processing using software in a fast manner.
In order to cover as many space domains as possible, we performed an analysis of different classes of on-board algorithms and we are currently designing a benchmarking suite for the evaluation of both embedded GPUs as well as their programming models. The benchmark suite includes applications such as image pre-processing, standard compression, FFT, FIR, and other common on-board processing tasks.
While our analysis in the algorithm selection points out that most of the image-related processing algorithms used in observation systems are a good fit for embedded GPUs, we have identified that the compression algorithms are among the most challenging ones. In this paper, we present our experience

with the GPU parallelisation of several components from the CCSDS Compressors 121, 122 and 123, which are included in our upcoming benchmark suite. In particular, we implement the following parts of each compression standard:

- CCSDS 121.0-B-2
  - Predictor
    - Unit-delay
    - error mapper
  - Encoder
    - Fundamental sequence
    - Sample Split
    - Zero block
- CCSDS 122.0-B-2
  - 2D multilevel wavelet transform
  - Bit planar encoder
- CCSDS 123.0-B-1
  - Predictor:
    - Adaptive weighted predictor
  - Encoder:
    - Simple and Block adaptive encoder

In particular, we focus only on the parts of the standards related to the encoding (e.g. forward transforms), since this part is expected to be used on-board. Moreover, we prioritise the parts which are a better fit for GPU parallelisation based on our analysis of their access patterns. In a future edition of the benchmark suite, we may consider to gradually add the non-implemented parts of the CCSDS.

The algorithms are implemented in multiple parallel programming models GPUs (CUDA, OpenCL), which is our primary focus. In addition, implementations have been made for CPUs (OpenMP) to exploit also the multicore capabilities of existing COTS SoCs featuring embedded GPUs. In our paper, which is an early preview of a significant part of our benchmarking suite, we will describe in detail the approach we have followed for each algorithm and provide insights about the different parallelisation approaches for GPUs and CPUs. We will present results with several state-of-the-art embedded GPU platforms, which have been selected as good candidate platforms earlier in the project [2], including the latest NVIDIA platform, Xavier, showing the performance benefits provided by the use of embedded GPUs. One major challenge for the use of GPUs in space is the requirement for fault-tolerance. Hence, we have targeted also the benchmarks on a fault-tolerant model based on a COTS GPU.

This work has been performed in the framework of the ESA GPU4S (GPU for Space) Project ITT AO/1-9010/17/NL/AF "Low Power GPU Solutions for High Performance On-Board Data Processing".

References:

[1] Leonidas Kosmidis, Jérôme Lachaize, Jaume Abella, Olivier Notebaert, Francisco J. Cazorla, David Steenari. GPU4S: Embedded GPUs in Space, 22nd Euromicro Conference on Digital System Design (DSD), 2019

[2] Leonidas Kosmidis, Iván Rodriguez, Álvaro Jover, Sergi Alcaide, Jérôme Lachaize, Jaume Abella, Olivier Notebaert, Francisco J. Cazorla, David Steenari. GPU4S: Embedded GPUs in Space - Latest Project Updates, Elsevier Microprocessors and Microsystems, Volume 77, September 2020

[3] Iván Rodriguez, Leonidas Kosmidis, Olivier Notebaert, Francisco J. Cazorla, David Steenari. An On-board Algorithm Implementation on an Embedded GPU: a Space Case Study, Design, Automation & Test in Europe Conference & Exhibition (DATE), 2020

*************************************************************

## Parallelization of Prediction and Encoding for Multispectral and Hyperspectral Images

Mr. Ioan Cristian Trohin[1], Mr. Lucian Banu[1], Ms. Roxana Cornelia Andrada Tomescu[1], Mr. Cosmin Padureadu[1]
*[1]Enea Services Romania S.R.L., Bucharest, Romania*

Parallelization of Prediction and Encoding for Multispectral and Hyperspectral Images
Enea Services Romania S.R.L.

This presentation analyses the parallelization schemas used for the prediction and encoding phases, described by the CCSDS 123.0-B-1 - Lossless Multispectral & Hyperspectral Image Compression and CCSDS-123.0-B-2 - Low-Complexity Lossless and Near-Lossless Multispectral and Hyperspectral Image Compression standards.

The CCSDS 123.0-B-1 standard is applicable for lossless compression, where the samples in each band are arranged in raster-scan order and each spectral band has its own weight vector, which is maintained independently of the weight vectors used by other spectral bands. Because of this, the parallelization method used for the prediction phase, is to divide the image along the z and y axes, in blocks made up of samples from entire x axis.

In comparison with the CCSDS 123.0-B-1, for the prediction phase from the CCSDS-123.0-B-2 standard, the same parallelization scheme could not be applied, because the quantization concept is introduced - which leads to some dependencies between consecutive spectral bands: the value computed for a given index in the current spectral band, depends by the value, from the same index, computed on the previous bands. These leads to a new parallelization scheme, where the image is divided along the z axis, in blocks made up of samples from entire y and x axis. In case the spectral bands are lower than the number of cores used, the image could be divided along the y axis and the blocks will be made up of samples from entire z and x axes.

Regarding the encoding phase, there are 3 encoding methods: Sample Adaptive and Block Adaptive defined in CCSDS 123.0-B-1 and CCSDS-123.0-B-2 and Hybrid method, which is a new concept introduced in CCSDS-123.0-B-2.

The scheme proposed when mapped prediction residuals are encoded using the Sample-Adaptive and the variable length code-words are saved in BI order, is to is partition the mapped prediction residuals along the y and the z axes in segments of data that can be encoded independently. The partition size along the z axis is given by the interleaving depth (M), except for, possibly, the last partition; while the partition size along the y axis is 1, which means that only one row is processed at a time. The sub-frame interleaving depth (M), when band-interleaved encoding order is used, could be selected so that the processing load on each core is balanced.

In case the variable length code-words are saved in BSQ order and the encoding method is Sample-Adaptive, the parallelization scheme is to divide the mapped prediction residuals along the z-axis in blocks of data that can be encoded independently. The size of each partition taken along the z-axis is calculated so that the task load can be managed efficiently. This method of distributing the computational load across processor's cores is facilitated by the fact that separated statistics (based on accumulator Σz(t) and counter Γ(t) values) are maintained independently for each spectral band.

The parallelization scheme chosen for Block-Adaptive has similarities to the one described for the Sample-Adaptive, the difference consists in the way in which the size of each partition is calculated. For the Block-Adaptive encoding method, the size of each partition is calculated based on the number of reference blocks that exist in the image, while for the Sample-Adaptive method the size of each partition is calculated based on the number of spectral bands.

The Hybrid encoding method can be described as a combination between Sample-Adaptive and Block-Adaptive, where 16 low-entropy codes are used alongside high entropy encoding. The Hybrid encoding will be performed in two stages: the low entropy code detection and the codewords detection and encoding. Due to this, the parallelization will consist in one processing block for each low entropy code, and one additional for the high entropy mapped quantizer indexes - that gives the chance to balance the load evenly among cores.

*********************************************************

## Implementation of cloud detection and processing algorithms and CCSDS-compliant hyperspectral image compression for CHIME mission

Yubal Barrios[2], Eng. Pedro Rodríguez[1], Antonio Sánchez[2], Roberto Sarmiento[2], Luis Rafael Berrojo[1], María Isabel González[1]

*[1]Thales Alenia Space In Spain, Tres Cantos, Spain, [2]Institude for Applied Microelectronics, University of Las Palmas de Gran Canaria, Las Palmas, Spain*

Hyperspectral sensors are increasing their presence on-board satellites because they provide relevant information for the scientific community in many applications. This is the reason why ESA has included the Copernicus Hyperspectral Imaging Mission for Environment (CHIME) in the future Copernicus 2.0 program. CHIME shall provide contiguous spectral coverage in VNIR and SWIR spectral domain (covering approximately 220 bands between 400nm and 2500nm). Acquisitions shall have typically 106Msps, with a dynamic range of 16 bits per sample and acquired in Band-Interleaved by Line (BIL) order, what leads to input data rates up to almost 2Gbps. However, clouds are estimated to cover more than 54% of the Earth's land area and 68% of the oceans. Many scientific applications which need to estimate Earth surface properties from satellite images are useless in presence of clouds, making more than half of the acquired scenes unusable. At sensor level, on board detection of cloudy areas and compression combined are then proposed to transmit sensor information to ground within a restricted time and with a limited downlink bandwidth.

This work, done in CHIME (phase A/B1) framework, presents a demonstrator including implementation of cloud detection and processing algorithms and hyperspectral image compressor based on CCSDS 123.0-B-2 standard, recommended by the Consultative Committee for Space Data Systems.

Cloud algorithms are composed by two stages: detection and processing. The cloud detection is performed by a Support Vector Machine approach (SVM). The SVM algorithm is pixel-based and it is followed by a simple filtering in order to reduce the false positive detection. The output is a spatial mask (with same number of pixels than the image) indicating for each hyperspectral pixel if it is cloud or not.

For the pixels detected as cloud, a pre-quantization is done, to improve the posterior compression in these less useful areas. Even for cloudy zones, some bands of the image can have scientific or commercial utility, therefore some interesting selected bands can be excluded from the processing.

The CCSDS-123.0-B-2 standard defines a lossless and near-lossless compression solution that specifically targets multispectral and hyperspectral images. The near-lossless compression mode is achieved by introducing a quantization loop in the prediction architecture of its predecessor, the CCSDS-123.0-B-1 lossless standard, and by controlling the compression losses through user defined error values.

To adapt the CCSDS-123.0-B-2 standard to CHIME mission requirements a tuning was performed, in order to identify the most suitable values for the different compression parameters proposed in the standard. With a set of images generated from AVIRIS images, to be representative of CHIME scenario, a wide number of simulations were done in order to fine tune parameters such as: Prediction mode, type of Local Sum, Number of Prediction Bands (P), Sample representative resolution (ϴ), Sample representative damping (ϕz), etc. As entropy coder, the CCSDS 121.0-B-2 block-adaptive coder was selected as a trade-off between compression capability and complexity.

Cloud algorithms (detection and processing) modules have been coded by means of RTL VHDL description. In case of the CCSDS-123.0-B-2 standard, it was modeled in C language and implemented by using High Level Synthesis (HLS) techniques, that automatically generate the equivalent RTL description.

The system has been successfully validated over the Xilinx KCU105 evaluation board, that mounts a Xilinx KU040 FPGA, new generation device representative of flight hardware. The design consumes low resources leaving growth potential for further evolutions or other additional applications.

This paper will present the design implementation, the demonstrator set-up and the validation plan to verify the correct behavior and performances. The test procedure consists in iterations between the simulations of the VHDL in workstation and the demonstrator hardware tests. The compressed images have been validated by comparison with references generated by the external CNES software, compliant with the compression standard. Several test vectors have been defined for the validation with different features, such as image size or percentage of clouds in the scene.

Acknowledgments: The research leading to these results has been financed by ESA in the frame of the CHIME phase A/B1 pre-developments. The authors want to thank Mickael Bruno and Mathieu Albinet from CNES for the compression software. Also thanks to Roberto Camarero and Raffaele Vitulli from ESA and Michel-François Foulon and Dimitri Lebedeff from Thales Alenia Space in France for their contributions.

*********************************************************

## Session 3 – Compression Algorithms

### On-board cloud detection and selective spatial/spectral compression based on CCSDS 123.0-B-2 for hyperspectral missions

Mr. Dimitri Lebedeff1, Mr Michel François Foulon1, Mr Roberto Camarero2, Mr Raffaele Vitulli2, Mr Yves Bobichon1

*1Thales Alenia Space, Cannes La Bocca / Toulouse, France, 2European Space Agency , Noordwijk, The Netherlands*

Clouds are estimated to cover about 67% of Earth surface, and even after elimination of partially cloudy and cloud edge pixels, the total cloud cover remains close to 50%. Many applications which need to estimate Earth surface properties are useless in the presence of such opaque clouds, making half of the acquisitions unusable for Earth science applications.

Future hyperspectral satellite missions will provide numerous bands in VNIR and SWIR domain, with large swath and small spatial sampling distance. Therefore, the amount of data to be transmitted to ground is large, and image compression becomes mandatory.

CCSDS SLS-MHDC Working Group has established a recommended standard for a low-complexity data compression applied to multispectral and hyperspectral sensors. It provides an effective method of performing lossless or near-lossless compression, with a control of the error introduced, essential for scientific applications. The error can be band-dependent, however, the standard does not include a possibility for a selective compression in the spatial dimension.

Considering the significant presence of clouds, in particular for missions with continuous Earth acquisitions, the aim of this work, is to explore some possibilities to increase the hyperspectral compression performance on-board of satellite with a selective compression applied to clouds.

The first step of the compression scheme is the cloud detection. Cloud detection is widely used on-ground for cloud classification. Several methods are operational, such as the physical approach (or “threshold” approach) applied to Landsat and to Sentinel-2, or the approach based on Support Vector Machine (SVM) applied for example on a French commercial program. Some experiments have also been performed for satellite on-board detection with threshold approach.

For on-board compression purpose, only opaque clouds that hide the ground need to be detected. The threshold and the SVM approaches have been defined according to this need, and performance assessed on Landsat cloud data base. The SVM approach has been selected as it presents a high adaptability to evolutions and is simple to implement when based on well-chosen spectral bands. The SVM parameters are expected to be stable, and are defined on-ground, thanks to a training stage and to a consistent cloud data base. The SVM cloud detection is pixel-based, but it is followed by a simple filtering that requires few lines in order to reduce the false positive detection. The output of cloud detection is a spatial map, which identifies each hyperspectral pixel as a binary “cloud” or “ground”.

The principle of selective compression is to adapt the compression to the different classes. In the case of cloud compression, the objective is to apply a higher loss on the pixels detected as cloud compared to the ground pixels for an improved data rate reduction. Three different approaches have been studied, all based on the Multispectral and Hyperspectral image compression CCSDS standard. The first one applies a pre-quantization to the cloud pixels before getting into the compressor, which is still fully standard-compliant. The two other approaches occur inside the CCSDS compression - one by an adaptation of the prediction stage to the two classes of pixels - the other one by directly operating on the output of the prediction, thus incorporating non-standard features to support differentiated compression settings based on pixel class.

The performances of the three approaches have been assessed on hyperspectral AVIRIS images and on one simulated scene representative of the future Copernicus Hyperspectral Imaging Mission (CHIME).

Implementation complexity has been assessed by Thales Alenia Space in Spain and University of Las Palmas de Gran Canaria, and is subject to another abstract submission.

This paper will give an overview of the cloud compression scheme, and will focus on the results obtained on the SVM cloud detection and on the cloud compression. Data rate reduction and impact on cloudy pixels after decompression will be presented together with some additional considerations to help for a selection of cloud compression scheme.

The study has been financed by ESA in the frame of CHIME phase A/B1 pre-developments.
The authors thank Manuel Sanchez from Thales Alenia Space Spain, Roberto Sarmiento, Yubal Barrios, Antonio Sánchez from University of Las Palmas de Gran Canaria for software development of CCSDS with evolutions and fruitful discussions, Mickael Bruno and Mathieu Albinet from CNES for CCSDS 123.0-B2 software used as reference.

**********************************************************

## Improving Storage Size and Random Access Time in k²-raster Compact Data Structure for Hyperspectral Scenes

Mr. Kevin Chow[1], Mr. Dion Eustathios Olivier Tzamarias[1], Dr. Ian Blanes[1], Dr. Joan Serra-Sagristà[1]
*[1]Universitat Autònoma de Barcelona, Cerdanyola del Valles, Spain*

Compact data structures are a type of losslessly compressed data structures that provide efficient storage and real-time processing. The data can be loaded into memory and accessed directly using the rank and select functions available from these structures. They facilitate better transmission through communication channels with limited bandwidth and provide faster data access. Queries to their data elements can be done without any need for full decompression.

Hyperspectral data are scenes taken from instruments on board an aircraft such as AVIRIS (Airborne Visible/Infrared Imaging Spectrometer), or on a satellite in space such as Hyperion and IASI (Infrared Atmospheric Sounding Interferometer). These scenes are made up of multiple bands from across the electromagnetic spectrum. Data extracted from these bands are useful in many diverse applications, for example, oil field search, mineral search, weather prediction, and wildfire soil studies, to name just a few.

Hyperspectral scenes are usually compressed to reduce their large sizes in order to facilitate their transmission.

In our research, we have been using a compact data structure called k²-raster, proposed by Ladra et al. [1], to compress hyperspectral data. The k²-raster was developed from k²-tree, another compact data structure used for web graphs and social networks. The difference between the two is that k²-raster is built from a matrix with integers while k²-tree is from a bit matrix. In our previous papers [2,3], we presented some of the experimental results that show that k²-raster can help reduce the size of hyperspectral data to as much as 52%. Another major advantage of this structure is its ability to randomly access data without full decompression, saving access time. DACs used in the original paper of k²-raster [1] is the only integer encoder that has been examined to allow the structure to do that. But as explained below, we show that other integer encoders also have this random-access capability with competitive results and therefore can be used as a substitute for DACs.

In this paper, we describe some of the improvements we made on two important aspects of k²-raster: data storage and access speed. First for storage, before the k²-raster tree is built, the original matrix may need to be extended in size so that it can be partitioned into equal-sized subquadrants. Doing so will facilitate the search for child nodes. However, this also creates a lot of empty nodes filled with zeros and the k²-raster results in a larger, sometimes bloated, storage size. To resolve this, we examined a method where the original matrix size is used to build the k²-raster tree without any expansion. The paper describes the steps to do so and shows some experimental results.

Another improvement is to find an alternative to the DACs integer encoder [4] which is used together with k²-raster to provide fast random access. We have been studying the possibility of using word-aligned integer encoders such as Simple9 [5], Simple16 [6] and PForDelta [7] to attain a better storage size and random-access time. The results are reported and compared to those of DACs.

Bibliography:
[1] Ladra, S., Paramá, J. R., & Silva-Coira, F. (2017). Scalable and queryable compressed storage structure for raster data. Information Systems, 72, 179-204.
[2] Chow, K., Tzamarias, D. E. O., Blanes, I., & Serra-Sagristà, J. (2019). Using Predictive and Differential Methods with K²-Raster Compact Data Structure for Hyperspectral Image Lossless Compression. Remote Sensing, 11(21), 2461.
[3] Chow, K., Tzamarias, D. E. O., Hernández-Cabronero, M., Blanes, I., & Serra-Sagristà, J. (2020). Analysis of Variable-Length Codes for Integer Encoding in Hyperspectral Data Compression with the k²-Raster Compact Data Structure. Remote Sensing, 12(12), 1983.
[4] Brisaboa, N. R., Ladra, S., & Navarro, G. (2013). DACs: Bringing direct access to variable-length codes. Information Processing & Management, 49(1), 392-404.
[5] Anh, V. N., & Moffat, A. (2005). Inverted index compression using word-aligned binary codes. Information Retrieval, 8(1), 151-166.
[6] Zhang, J., Long, X., & Suel, T. (2008, April). Performance of compressed inverted list caching in search engines. In Proceedings of the 17th international conference on World Wide Web (pp. 387-396).
[7] Zukowski, M., Heman, S., Nes, N., & Boncz, P. (2006, April). Super-scalar RAM-CPU cache compression. In 22nd International Conference on Data Engineering (ICDE'06) (pp. 59-59). IEEE.

**********************************************************

## CILLIC: Context Interpolation Lossless and Lossy Image Compressor

Dr. Jordi Portell[1,2], Mr. Riccardo Iudica[1], Dr. Alberto G. Villafranca[1,3], Miguel Hernández-Cabronero[4], Ian Blanes[4], Joan Serra-Sagristà[4]
*[1]DAPCOM Data Services S.L., Vic, Spain, [2]Dept. FQA, Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Barcelona, Spain, [3]STAR-Barcelona S.L., Sant Cugat del Vallès, Spain, [4]Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain*

Image data compression requires a spatial decorrelation algorithm and, in case of multi/hyperspectral images, a spectral decorrelator. In satellite payloads, spatial decorrelation is often based on transforms such as DWT [1]. Spectral decorrelators can use transforms like POT [2,3], although prediction-based algorithms can be faster and compress even better [4]. Many of these solutions are often designed with hardware implementations in mind, which can make their software equivalents difficult and with high computing and memory requirements.

We have designed CILLIC, a new image data compression algorithm. This block-based spatial and spectral decorrelator has been designed with software implementations in mind, aiming at an efficient use of CPU caches, minimizing loops and conditions, and reducing memory requirements. It is inspired by HPA [5], also working on the spatial domain and using simple integer operations without any transform.

CILLIC works in blocks of 15x15 pixels and uses different decorrelators: spatial, spectral and mixed. The spatial decorrelator defines different pixel types within the block, allowing to predict most of the pixels from the interpolation of their neighbours. More than half of the pixels are predicted from the interpolation of four neighbours and one eighth from two neighbours. We use this spatial approach in single-band images and in the first band of multi/hyperspectral images. The spectral decorrelator predicts each pixel from the co-located pixel of the previous spectral band plus an average inter-band difference. Finally, the mixed decorrelator, based on the multi-band predictor of FAPEC [6], uses a linear combination of spatial and spectral neighbours to predict each pixel.

In multi-band images, from the second band onwards, CILLIC performs a quick trial of each decorrelator on a small subset of each block and selects the one leading to the smallest residuals. The option chosen is explicitly flagged, allowing to fine-tune the algorithm (or increase the size of the testing subset) without requiring changes in the decompressor. This approach allows for a better adaptation to artefacts in uncalibrated multi/hyperspectral images, either in the spatial or spectral domain, such as streaks, overexposed or underexposed bands.

Near-lossless operation (fixed-quality) is implemented by using different quantization levels on the residuals. Some reference pixels on each block use conservative quantization levels, whereas the rest use more aggressive quantization. Owing to the spatial interpolation algorithm, this approach keeps most of the sharp image features even at moderate loss levels. The average variation in brightness due to losses is determined for each block and explicitly flagged, allowing for an unbiased reconstruction even at high loss levels. Each pixel is reconstructed during compression before being used in subsequent calculations, avoiding the propagation of losses through blocks or bands.

Lossy operation (fixed-rate) has also been implemented. The actual compression ratio is periodically estimated and compared against the target ratio. Then, the loss level is updated, if necessary, to converge towards the target ratio.

CILLIC is implemented as a new pre-processing stage of FAPEC, taking advantage of its highly optimized and versatile software framework, as well as of its efficient and outlier-resilient entropy coding core. Our tests reveal that CILLIC offers ratios and PSNR levels very similar (sometimes better) than DWT, whereas its execution speed can be much faster (more than twice in some cases, specially in lossless). The overall design of FAPEC and CILLIC makes them suitable even for low-end computing platforms, such

as on-board computers or cubesats. The versatility of this solution makes it applicable to demanding applications such as real-time in-orbit video compression on high-end computing platforms.
In this work we present this new algorithm and we compare its lossless, near-lossless and lossy performances against other solutions, such as DWT and HPA of FAPEC, CCSDS 122.1-B-1 and CCSDS 123.0-B-2.

[1] Image Data Compression (2017), https://public.ccsds.org/Pubs/122x0b2.pdf
[2] I. Blanes, J. Serra-Sagristà (2011) "Pairwise Orthogonal Transform for Spectral Image Coding," IEEE TGRS 49:3 961-972, 10.1109/TGRS.2010.2071880
[3] Spectral Pre-processing Transform for Multispectral and Hyperspectral Image Compression (2017), https://public.ccsds.org/Pubs/122x1b1.pdf
[4] Low-Complexity Lossless and Near-Lossless Multispectral and Hyperspectral Image Compression (2019), https://public.ccsds.org/Pubs/123x0b2c1.pdf
[5] R. Iudica, J. Portell, E. García-Berro (2014) Hierarchical Pixel Averaging: A New Image Compression Approach, in OBPDC IV, ESA/CNES.
[6] J. Portell, R. Iudica, E. García-Berro et al. (2017) FAPEC, a versatile and efficient data compressor for space missions, IJRS 39:7, 2022-2042, 10.1080/01431161.2017.1399478

*********************************************************

## FASEC: Fast and Simple Entropy Coder

Dr. Jordi Portell[1,2,3], Mr. Màrius Montón[3], Mr. Riccardo Iudica[1], Dr. Alberto G. Villafranca[1,4]
*[1]DAPCOM Data Services S.L., Vic, Spain, [2]Dept. FQA, Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Barcelona, Spain, [3]Institut d'Estudis Espacials de Catalunya (IEEC), Barcelona, Spain, [4]STAR-Barcelona S.L., Sant Cugat del Vallès, Spain*

FAPEC [1], the Fully Adaptive Prediction Error Coder, is a high-performance and versatile data compressor which encompasses a suite of pre-processing algorithms for images, time series or even text files. Tailored stages can be implemented as well, to better adapt to specific kinds of data. It relies on PEC [2], an outlier-resilient entropy coding core with run-length and low-entropy extensions. FAPEC is, currently, one of the best options for cubesats, as demonstrated by the Spire Global constellation already using FAPEC since late 2017. However, despite its excellent performance, some scenarios may require even faster solutions with even smaller code and memory footprints.
On-Board Computers (OBC), not to be confused with On-Board Data Handling systems (OBDH), are a clear example of this. OBCs must be extremely reliable to guarantee the health of the satellite and to ensure its correct operation. Thus, they are typically based on processors, Systems-on-Chip (SoC) or microcontrollers with very high reliability, such as redundant or even space-qualified components. Some sort of real-time operating system (RTOS) is often used, imposing significant restrictions on the software implementation, such as the impossibility to use dynamic memory allocation. Computing power, memory and storage are extremely limited in OBCs, as well as downlink bandwidth, since it is mainly intended for housekeeping and essential telemetry. All these limitations are even worse for OBCs in cubesats and small satellites.
On the other hand, for simplicity, reliability or simply due to cost limitations, some cubesats and small satellites only have an OBC, without any OBDH that could provide more computing and downlink resources. This poses very tight restrictions on the amount of data that can be acquired and transmitted. Data compression could obviously be a solution to this, but the extreme limitations mentioned are a challenge difficult to overcome. Hardware-based data compression could solve this, but it would require either expensive dedicated chips or a programmable hardware device (such as an FPGA), resulting in

higher complexity and development costs. Either option is quite blocking in the NewSpace paradigm, which aims at agile solutions typically based on software. In the end, even a very modest (but extremely fast) data compressor can provide a significant added value to an OBC-only mission, because in nearly all cases the alternative is not to compress the data.

In this work we present an alternative entropy coding core for the FAPEC data compression framework, which we have called FASEC (Fast and Simple Entropy Coder). FASEC has been designed to minimize condition checks and loops, which are some of the bottlenecks in the original FAPEC core. Bytewise memory accesses are enforced, instead of bitwise ones, to further optimize its operation. As a proof of concept, the current implementation only supports 8-bit samples. Only a simple pre-processing stage has been embedded in the same entropy coder, namely, delta decorrelator with or without interleaving. Together with some optimizations done to the original FAPEC framework, we have been able to run it under FreeRTOS on a Cortex-M4 160 MHz SoC with only 128 KB of RAM and 512 KB of flash memory, developed by IEEC as part of its C3SatP cubesat platform [3]. Preliminary tests reveal that FASEC can double the throughput of FAPEC at the cost of a moderate reduction in the compression ratio, as otherwise expected. Here we present these tests, which demonstrate the feasibility of running a data compressor on an OBC, thus boosting its downlink capabilities without compromising reliability. Future developments will include support for 16-bit samples, further optimizations, and the integration of FASEC in more complex pre-processing stages.

[1] J. Portell, R. Iudica, E. García-Berro, A. G. Villafranca and G. Artigues (2017) FAPEC, a versatile and efficient data compressor for space missions, International Journal of Remote Sensing, 39:7, 2022-2042, DOI: 10.1080/01431161.2017.1399478

[2] J. Portell, E. García-Berro and A. G. Villafranca (2010) Quick outlier-resilient entropy coder for space missions, Journal of Applied Remote Sensing 4(1), 339-363, DOI: 10.1117/1.3479585

[3] J. Ramos-Castro, J. Colomé, J.M. Gómez-Cama et al. (2019) High-performance on-board computer and comms for cubesats, in European Workshop on On-Board Data Processing 2019, edited by ESA/CNES/DLR, ESTEC, The Netherlands, February 25-27

*************************************************************

### Satellite Constellation Data Compression

Mr. Ashwin Kumar Gururajan[1], Dr Miguel Hernandez-Cabronero[1], Dr Ian Blanes[1], Dr Joan Serra-Sagrista[1]
*[1]Universitat Autonoma de Barcelona, Barcelona, Spain*

Small satellites deployed as a network in space have a significant advantage over conventional satellites because of their low entry barrier and the inherent capability of distributed multi-node systems. Several small satellites flying in a formation would collaboratively achieve the mission aim at lower cost and with enhanced reliability and efficiency compared to larger single platform mission. There is heavy research in the field of satellite data compression wherein, data is compressed onboard the satellite and then sent to the ground station for processing, while research on satellite constellation systems where we could further exploit the properties of multiple close satellites to achieve higher compression with the help of inter-satellite links is just beginning. This paper explores this topic further and we analyse how Earth Observation (EO) satellites in a low earth orbit (LEO) could exploit these properties further for higher compression gains over existing methods.

*************************************************************

## Session 4 – Innovative Data Compression & Reduction Techniques

### Validated efficient image compression for quantitative and AI applications.

Dr. Bruno Sanguinetti[1], Dr. Christoph Clausen[1], Mr. Michael Desert[1], Ms Evgeniya Balysheva[1]
*[1]Dotphoton Ag, Zug, Switzerland*

An increasing proportion of earth observation applications use quantitative and AI algorithms. In this paper we consider the requirements that such algorithms impose on image compression and how to validate them. We then present our approach to raw optical image compression which achieves a compression ratio  in the range 5:1–10:1 with a SNR loss 1.25dB at up to 200Mpix per second in both software and FPGA.

Historically, lossless image compression has been used to handle raw data, as it allow for accurate post-processing, easy archival and translation across different lossless formats.  This comes at a price of a low compression ratio, typically 1.5:1 an rarely >2:1. Higher compression ratios can be achieved at the expense of image information loss and limited flexibility, as chaining lossy compression algorithms may generate unforeseen interactions and artefacts. For consumer AI applications, these types of lossy compression is acceptable, however to achieve their high compression ratios, they often sacrifice the fine information which allows AI algorithms to achieve image enhancements such as sharpening, superresolution, denoising, segmentation and many others in a qunatitatively accurate way. To ensure that a compression algorithm is suitable for both foreseen and unforeseen applications, a general design and testing approach can be taken that yields a clear specification on the suitability of images at a given compression ratio. The approach that we present here relies on a physical model of the sensor with which the image was taken and generates a compressed raw image which is statistically consistent with the image arising from another (or the same) sensor of given specifications, and then being losslessly compressed and therefore suitable for post-processing, archival or format translation.
In particular, we demonstrate how these concepts can yield an algorithm able to provide the high performance stated above with constrained power and FPGA footrint.

We then discuss the advantages of embedding the physical sensor model within the data in the context of AI. A first use consists in data normalization by mapping sereval sensor models to a single output model therefore enhancing training.  A second example is the reverse:  data augmentation by simulating the statistics arising from different sensors. A third example is monte-carlo uncertainty propagation which gives a rapid overview of the uncertainty for even a complex AI algorithm.

We give an outlook by imaginig what a future AI image-processing pipeline could look like from the data management point of view, taking into account metrological accuracy, performance, cost and practicality for the end-user.

*********************************************************

### Cots Based Electronic Video Chain With Advanced Digital Processing For High Resolution Earth Observation Satellites

Dr. Jean-Pierre Millerioux[1], Alex Materne[1], Dr Carole Thiebaut[1], Laurent Lebegue[1], Sophie  Petit[1], Florie Languille[1], Lionel Perret[1], Laurent Boukris[2], Ursula Kirchgaessner[2], Sylvain Angebault[2]
*[1]CNES, Toulouse, France, [2]Nexvision, Marseille, France*

This paper describes CNES activities around a new concept of electronic video chain for high resolution and low cost optical Earth Observation Satellites. This concept is based on the use of COTS CMOS sensors and a Xilinx Ultrascale FPGA. A breadboard of the electronic video chain, called ‘DARWIN-CU’, is realized to proof the concept and achieve TRL 5. The overal system design is presented and discussed.

**********************************************************

### The SURPRISE demonstrator: a super-resolved compressive instrument in the visible and medium infrared for Earth Observation

Dr. Valentina Raimondi, Luigi Acampora[1], Gabriele Amato[1], Massimo Baldi[1], Dirk Berndt[7], Alberto Bianchi[8], Tiziano Bianchi[3], Valentina Colcelli[1], Chiara Corti[5], Francesco Corti[5], Marco Corti[5], Nick Cox[4], Ulrike Dauderstädt[7], Peter Dürr[7], Sara Francés Gonzáles[7], Paolo Frosini[6], Donatella Guzzi[1], Jessica Huntingford[6], Demetrio Labate[8], Nicolas Lamquin[4], Cinzia Lastri[1], Enrico Magli[3], Vanni Nardino[1], Lorenzo Palombi[1], Irene Pettinelli[6], Giuseppe Pilato[8], Alexandre Pollini[2], Enrico Suetta[8], Diego Valsesia[3], Michael Wagner[7]

*[1]CNR – IFAC, Sesto Fiorentino, Italy, [2]Centre Suisse d'Electronique et Microtechnique, Neuchâtel, Switzerland, [3]Politecnico di Torino - DET, Torino, Italy, [4]ACRI-ST, Sophia-Antipolis, France, [5]SAITEC srl, Firenze, Italy, [6]RESOLVO srl, Firenze, Italy, [7]IPMS – Fraunhofer Institut, Dresden , Germany, [8]LEONARDO S.p.A., Campi Bisenzio , Italy*

While Earth Observation (EO) data has become ever more vital to understanding the planet and addressing societal challenges, applications are still limited by revisit time and spatial resolution. Though low Earth orbit missions can achieve resolutions better than 100 m, their revisit time typically stands at several days, limiting capacity to monitor dynamic events. Geostationary (GEO) missions instead typically provide data on an hour-basis but with spatial resolution limited to 1 km, which is insufficient to understand local phenomena.

In this paper, we present the SURPRISE project - recently funded in the frame of the H2020 programme – that gathers the expertise from eight partners across Europe to implement a demonstrator of a super-spectral EO payload - working in the visible (VIS) - Near Infrared (NIR) and in the Medium InfraRed (MIR) and conceived to operate from GEO platform -with enhanced performance in terms of at-ground spatial resolution, and featuring innovative on-board data processing and encryption functionalities. This goal will be achieved by using Compressive Sensing (CS) technology implemented via Spatial Light Modulators (SLM). SLM-based CS technology will be used to devise a super-resolution configuration that will be exploited to increase the at-ground spatial resolution of the payload, without increasing the number of detector’s sensing elements at the image plane. The CS approach will offer further advantages for handling large amounts of data, as is the case of superspectral payloads with wide spectral and spatial coverage. It will enable fast on-board processing of acquired data for information extraction, as well as native data encryption on top of native compression.

SURPRISE develops two disruptive technologies: Compressive Sensing (CS) and Spatial Light Modulator (SLM). CS optimises data acquisition (e.g. reduced storage and transmission bandwidth requirements) and enables novel onboard processing and encryption functionalities. SLM here implements the CS paradigm and achieves a super-resolution architecture. SLM technology, at the core of the CS architecture, is addressed by: reworking and testing off-the-shelf parts in relevant environment; developing roadmap for a European SLM, micromirror array-type, with electronics suitable for space qualification.

By introducing for the first time the concept of a payload with medium spatial resolution (few hundreds of meters) and near continuous revisit (hourly), SURPRISE can lead to a EO major breakthrough and complement existing operational services.

CS will address the challenge of large data collection, whilst onboard processing will improve timeliness, shortening time needed to extract information from images and possibly generate alarms. Impact is relevant to industrial competitiveness, with potential for market penetration of the demonstrator and its components.

*********************************************************

## Session 5 – AI for data reduction I

### Plenary speaker: Deep learning for satellite image processing: where are we going?

Enrico Magli[1]
*[1]Politecnico of Torino, Italy*

Artificial intelligence is playing a growing role in satellite image processing and in the remote sensing market, as deep learning methods are improving the performance of many processing and inference tasks. This talk will overview recent advances in the field, covering tasks related to image generation (e.g., image denoising/despeckling and super-resolution) as well as image classification, event prediction and compressive imaging. It will also discuss key aspects in the application of deep learning to onboard satellite image processing.

*********************************************************

### AIX smart processing services in orbit

Mr. Cristoforo Abbattista[1], Mr. Leonardo Amoruso[1], Mr. Stefano Antonetti[2], Mr. Lorenzo Feruglio[3]
*[1]Planetek Italia srl, Bari, Italy, [2]D-Orbit SpA, Fino Mornasco (CO), Italy, [3]AIKO srl , Torino, Italy*

Space mission's scenario is rapidly evolving. However, the transition from a traditional space model into a commercial one is already showing some bottlenecks and barriers in specific applications preventing new market opportunities to flourish and limiting the effectiveness of the services delivered to ground. In several verticals where EO data is today used (e.g. agriculture), users/customers are being focused on their core business process, and less interested in data itself and more in applications to make their business more efficient and effective.
In some context (e.g. fire detection) the right information shall be provided to end users/customers at the right time and in the right place. And the place can be in some cases the space segment, where the availability of actionable information can be a game changer. In this approach part of the EO value chain is moved from ground to space to promptly transform sensed data into "wisdom", to exploit it directly or to enable an optimized exploitation of limited resources on-board (in fire detection this means downlinking just the warning with its geographical location, few tens of bytes instead of raw GBs). We defined this approach as "SpaceStream".
Although some of the barriers are currently being addressed (e.g. launches availability and satellite deployment flexibility, versatile ground infrastructures and software, suitable regulations, etc.), existing satellite operators, incoming new satellite operators and end users of space data are still experiencing severe inefficiencies affecting traditional mission operational models:

- delays in decision-making due to ground operator intervention,

- missed observation opportunities due to limited on-board autonomy (data analysis is currently largely executed on ground),
- missed information due to limited downloadable data from satellites (small satellites usually offer limited power, bandwidth and controllability),
- poor quality or irrelevant downlinked data,
- late systems failures detection.

Autonomous operations and Artificial Intelligence on-board are key enabling technologies able to impact over these limitations. Main expected impacts will affect reactivity, responsiveness and latency.
The advanced functionalities based on detection, extraction and exploitation of data information content will be the core value points in this paradigm shift, moving focus beyond "raw" data. Most of these new functionalities will have to manage huge amounts of raw data, implement structured and automated information refinement processes, able to react in real or near real time to single pieces of information, to touch off new tasking for improved EO acquisitions specifically targeting the detected needs.

In order to provide such solutions to the highlighted problems, AI-express (AIx) aims at building a new AI enabled processing framework, embarking it on satellites and then bringing into the market a new concept of satellite as-a-service. AIx will make available in-orbit resources on-demand: EO data and above all, actionable information and actions at the right time.
AIx is a set of services à la carte provided by a flying satellite and based on a public catalogue with an "app store" approach. Services will include EO data acquisition, processing (actionable info extraction), downlink and distribution. Ready-made applications (e.g. a fire detection and warning service) will be made available on the app-store and services can be also combined together to build custom applications.
Useful information can be then transferred back to Earth as notifications and alerts, or directly exploited on-board in autonomous decisions workflows. Useless data and information can be completely deleted without loading any precious resource like memory and bandwidth.
At regime AIx services will be deployed on a fleet of AI enabled satellites (D-Orbit's ION enhanced platforms) and will be available to any user customer interested in testing applications based on AIx EO payloads' data on-board (and also to third-party payloads embarked on IONs). Commercial services will be based on the pay-per-use usage of the on-board assets that will be negotiated directly among on-board sub-systems and accounted thanks to blockchain type mechanisms.

What distinguishes AI-express from the rest of the market competition is the capability to provide a set of deeply configurable services, scaling from pay-per-use to full missions as-a-service. Services allow for the evaluation of new approaches to space missions and for the validation of novel concepts in the real environment. Flexibility and configurability allow for iterating in-orbit tests and fine tuning of applications.

**********************************************************

## On board images processing using IA to reduce data transmission: example of OpsSat cloud detection

Eng. Frédéric Feresin[1]
[1]*Institue Research & Technology, Valbonne, France*

Pending on flight results of OpsSat, we propose to present to you the long process to deploy Neural Networks on FPGA.

The first step, often underestimated with respect to its impact on the final performances, consist in collecting and labelling images to perform the NN learning. The dataset shall content good enough images to train the NN parameters (weights and bias) and large diversity to avoid overfitting. OpsSat camera has not flight experience so we have to plan an update of the NN once first pictures will be available to upgrade the parameters on the basis of “real” images from space of the sensor that is why the capability to upload new NN is necessary. This can be useful also to change or add a class detection during the life cycle depending on user needs.
Then the second step consist in selecting the NN architecture and size, depending on learning dataset quality and quantity, and depending on execution target (Integrated Circuit) capability. Considering the Cyclone V performances and the application selection, the clouds detection, we choose a Convolutional NN based on Lenet 5 architecture with some specific improvements. We implement also a very innovative approach based on "spikes" layers, in order to reduce the inference consumption.
Finally the third step is the deployment of the CNN on the Cyclone V FPGA. This last development step requires specific competences of RTL coding, cause not any commercial solution to deploy NN on Cyclone V are compatible to "small" FPGA. We apply a "pipelined" approach to optimize the execution time, considering the "tiny" CNN architecture is compatible to limited number of logic cells.
Thanks to this process, that we will detail in the final abstract, the on ground tests on MytiSOM board demonstrate only 150ms to infer a full OpsSat images (close to 2M pixels).
Hoping be able before the 7th OBPDC conference to present inferences performed on flight!

*********************************************************

## FPGA Acceleration of Quantised Neural Networks for Remote-Sensed Cloud Detection

Philippe Reiter[1], Philipp Karagiannakis[2], Murray Ireland[2], Steve Greenland[2], Dr. Louise Crockett[1]
*[1]University Of Strathclyde, Glasgow, United Kingdom, [2]Craft Prospect, Glasgow, United Kingdom*

The capture and transmission of remote-sensed imagery for Earth observation is both computationally and bandwidth expensive. In the analyses of remote-sensed imagery in the visual band, atmospheric cloud cover can obstruct up to two-thirds of observations, resulting in costly imagery being discarded [1]. Mission objectives and satellite operational details vary; however, assuming a cloud-free observation requirement, a doubling of useful data downlinked with an associated halving of delivery cost is possible through effective cloud detection. A minimal-resource, real-time inference neural network is ideally suited to perform automatic cloud detection, both for pre-processing captured images prior to transmission and preventing unnecessary images being taken by larger payload cameras.

Much of the hardware complexity of modern neural network implementations resides in high-precision floating-point calculation pipelines. In recent years, research has been conducted in identifying quantised, or low-integer precision equivalents to known deep learning models, which do not require the extensive resources of their floating-point, full-precision counterparts. Our work leverages existing research on binary, ternary and quantised neural networks (QNNs) to develop a real-time, remote-sensed cloud detection solution using a low-cost, commodity system-on-chip (SoC). This follows on developments of the Forward Looking Imager [2] for predictive cloud detection developed by Craft Prospect, a space engineering practice based in Glasgow, UK.

Recent neural network minimisation advances have examined reducing the number of layers and parameters in models to maintain a high degree of precision and accuracy at a fraction of the complexity. Even with hundreds of orders of magnitude reductions in network sizes, these neural nets still encompass hundreds of thousands of weights. At 32 bits per weight, the storage demands and computational stress from performing matrix calculations on these huge arrays of parameters continues to prevent novel networks from being implemented in real-time embedded devices.

To further reduce the computational load, weight quantisation is required. Thresholding each weight from a 32-bit floating-point value down to a fixed, 8-bit integer representation is commonly performed. Such a reduction in weight precision preserves the structure of an existing network and will minimally impact its inference accuracy. However, for resource-strapped and low-power devices, storing and computing 8-bit integer weights can remain too taxing. A more extreme method of quantisation is required in the form of 4-, 2- and 1-bit networks.

Field-programmable gate arrays (FPGAs) feature a number of advantages over application-specific embedded processors, predominantly the potential for creating custom instruction sets. They present opportunities for high-precision results while maintaining adequate performance levels and a hardware footprint amenable to the resource-restricted domain of remote sensing.

To achieve a highly parallel FPGA architecture, the open-source FINN framework, supported by Xilinx, is used for the synthesis of QNNs [3]. FINN efficiently supports the previously described optimisations, including variable quantisation. In addition, FINN can adjust designs for maximal usage of FPGA block memory – reducing latency and increasing system performance [4]. For real-time predictions, images are compressed ahead of FPGA processing via the general processing cores on the SoC. Our inference implementation uses a Xilinx Zynq-7000 SoC with integrated FPGA.

The primary dataset used for QNN training is a custom, annotated image repository using ESA Copernicus Sentinel-2 captures. A fully convolutional, binary neural network served as the benchmark for the deep learning pipeline. Ternary and low-integer networks were then compared to this baseline for performance, accuracy, and resource utilisation comparisons using roofline models across the heterogenous processing architecture.

References:

[1] Mohajerani, Sorour, Thomas A. Krammer, and Parvaneh Saeedi. “Cloud detection algorithm for remote sensing images using fully convolutional neural networks.” arXiv preprint arXiv:1810.05782 (2018).

[2] Greenland, Steve, Murray Ireland, Chisato Kobayashi, Peter Mendham, Mark Post, and David White. “Development of a minaturised forwards looking imager using deep learning for responsive operations.” ESA, 2018.

[3] Blott, Michaela, Thomas B. Preußer, Nicholas J. Fraser, Giulio Gambardella, Kenneth O’brien, Yaman Umuroglu, Miriam Leeser, and Kees Vissers. “FINN-R: An end-to-end deep-learning framework for fast exploration of quantized neural networks.” ACM Transactions on Reconfigurable Technology and Systems (TRETS) 11, no. 3 (2018): 1-23.

[4] Umuroglu, Yaman, Nicholas J. Fraser, Giulio Gambardella, Michaela Blott, Philip Leong, Magnus Jahre, and Kees Vissers. “Finn: A framework for fast, scalable binarized neural network inference.” In Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 65-74. 2017.

************************************************************

## Session 6 – AI for data reduction II

### The size matters: Onboard hyperspectral data reduction using deep learning

Dr. Jakub Nalepa[1], Mr Lukasz Tulczyjew[1], Mr Michal Myller[1], Dr. Michal Kawulok[1]
*[1]KP Labs, Gliwice, Poland*

Hyperspectral imaging can capture hundreds of images acquired for narrow and continuous spectral bands across the electromagnetic spectrum. Since the spectral profiles are specific for different materials, exploiting such high-dimensional data can help determine the characteristics of the objects of interest. A hyperspectral image can be interpreted as a data cube which couples spatial and spectral information captured for every pixel. Practical applications of such imagery are very vast and they spread across a number of fields, including, biology, medicine, forensics, and remote sensing.

The number of bands in hyperspectral images (HSI) can reach hundreds, and it is a very useful source of information in various remote sensing applications. However, its huge volume brings challenges in efficient analysis, transfer (especially sending the hyperspectral data acquired on board a satellite back to Earth is extremely time-consuming and costly), and storage of such imagery. Also, data redundancy is a serious practical issue. The neighboring bands are often correlated, therefore only their small subset contributes to the HSI classification process. Finally, generating ground-truth (manually-annotated) data for supervised classification and segmentation methods is extremely difficult, time-consuming, and prone to human errors, and exploiting small (in terms of the number of each-class examples) and very high-dimensional datasets can easily deteriorate the performance of supervised learners. To deal with these issues, HSI is subjected to either feature extraction (generating new, perhaps more informative, less redundant, and compressed features from HSI) or feature selection (determining a subset of all HSI bands which convey the most important and useful information). These techniques can drastically reduce the dimensionality of the original hyperspectral data, and – in ideal scenario – do not deteriorate the amount of information captured by such imagery.

In this talk, we review both feature extraction and band selection algorithms that benefit from deep learning. In the former case, deep learning-powered techniques elaborate new latent (and often significantly compressed) representations of the original data that can capture important information within the data. We focus on autoencoder-based deep network architectures, alongside various recurrent neural networks (long short-term memory- and gated recurrent unit-based models) that are applicable to both multi- and hyperspectral data, and show how they can be deployed in end-to-end data analysis pipelines. On the other hand, we review attention-based convolutional neural networks used for band selection, and show how understanding the influence of specific parts of the entire spectrum can help us select only a small subset of all bands that convey the most important information about the objects in the scene. Since the number of informative bands is often small compared to all available bands, the process of selecting “useful” bands may be considered as anomaly detection within all bands, according to the elaborated attention scores. We show how the dimensionality reduction of hyperspectral data affects all other steps in the hyperspectral image segmentation chain. Finally, as exploiting deep learning-powered techniques in hardware-constrained environments is challenging because of their memory and energy requirements, we perform quantization of the investigated models (to notably decrease their hardware requirements and make them more resource-frugal), and verify the influence of the quantization process on the overall quality of the feature extraction process. Our talk is concluded with a discussion on open questions and challenges in the state of the art that ultimately need to be tackled to allow us seamlessly deploy such deep learning-powered approaches on-board satellites.

*************************************************************

## Smart payloads : image analysis by deep learning on-board

Dr François De Vieilleville[1], Dr Adrien Lagrange[1], Dr. Rosario Ruiloba Quecedo[1], Ms Aurore Dupuis[2], M Stéphane May[2], M Jean-Pierre Millerioux[2], M Clément Coggiola[2], M Mickael Bruno[2]
*[1]Agenium Space, Toulouse, France, [2]CNES, Toulouse, France*

This paper presents first implementations of high performance Deep Neural Networks (DNN) embedded in FPGA material representative of the hardware available in new space missions. DNN inference on board is made possible by the definition and development of highly efficient simplification methods allowing to execute the best-in-class deep neural networks (DNN) with hundreds of millions of parameters in the limited processing resources available on-board.

Small satellites platforms expansion increases the need to simplify payloads and to optimize downlink data capabilities. A promising solution is to enhance on-board software, in order to take early decisions, automatically. However, the most efficient methods for data analysis are generally large deep neural networks (DNN) oversized to be loaded and processed on limited hardware capacities of small satellites. To use them, we must reduce the size of DNN while accommodating efficiency in terms of both accuracy and inference cost. In this paper, we present a distillation method which reduces image segmentation deep neural network's size to fit into on board processors. This method is presented through a ship detection example comparing accuracy and inference costs for several networks.

Distillation provides a way to extract the really meaningful parts of large and complex DNNs in a reduced model. This extraction is mainly performed by transferring the knowledge of a big teacher network in a smaller DNN by training the small DNN to predict the output of the teacher model. It shall not bring significant loss in terms of precision and reliability. Then, this size reduction, at no performance cost, will also simplify the inference code required to execute the distilled DNN on FPGA HW. This approach is complementary of existing techniques to reduce DNN memory footprints (ex. those implemented in EUCLID deep Space Mission, PhiSat-1 mission), compatible with precision reduction techniques for inference and can reuse generic VHDL (Very High Speed Integrated Circuit Hardware Description Language) code generation for running DNN on Soc FPGA HW. Our approach will reduce costs to fit state of the art DNN on HW of image payload.

Several DNN architectures have been implemented and simplified for ships segmentation and detection in very high resolution optical images. Simplified DNN were ported (inference code is adapted) and executed in mid-range FPGA HW (Xilinx ZCU102) without significant performance losses (< 10% F1-score). The implemented solutions and the obtained results (execution time on-board, frames per second processed, resources and power consumption) will be presented.

The work presented in this article was performed by AGENIUM Space and CNES in the framework of research and development contracts and internal activities aimed at the implementation of intelligent payloads.

*************************************************************

## Invited Speaker: CompressAI: A PyTorch library and evaluation platform for end-to-end compression research

Dr. Jean Bégaint[1], Dr. Fabien Racapé[1], Mr. Simon Feltman[1], Mr. Akshay Pushparaja[1]
*[1]Interdigital US, Palo Alto, United States*

This paper presents CompressAI, an open-source library that provides custom operations, layers, models and tools to research, develop, and evaluate end-to-end image and video codecs. In particular, CompressAI includes pre-trained models and evaluation tools to compare learned methods with traditional codecs. Multiple models from the state-of-the-art on learned end-to-end image compression have been reimplemented in PyTorch [1] and trained from scratch using the Vimeo90K2 training dataset [2].

The current deep learning ecosystem is mostly dominated by two frameworks: PyTorch and TensorFlow. Discussing the merits, advantages and features of one framework over another is beyond the scope of this document. There is evidence that PyTorch has seen major growth in the academic and industrial research circles over the last years. However, building end-to-end architectures for image and video compression from scratch in PyTorch requires a lot of re-implementation work, as PyTorch does not ship with any custom operations required for compression, such as entropy bottleneck and estimation tools. These tools are also mostly absent from the current PyTorch ecosystem, whereas the TensorFlow framework already has a compression library .

CompressAI aims to implement the above-mentioned operations to build deep neural network architectures for data compression in PyTorch and provide evaluation tools for comparing learned methods against traditional codecs. CompressAI includes custom layers, entropy models, operations, and models to build end-to-end codecs. CompressAI also provides pre-defined model architectures from the state-of-the-art [4]–[6], including pre-trained weights achieving similar performances as reported in the original papers. All the implemented models fully support end-to-end compression and decompression of images, with a bit-stream representation leveraging an entropy coder based on the range Asymmetric Numeral Systems algorithm [3]. Additionally, CompressAI provides some tools to facilitate the research on learned codecs. For example, the following traditional codecs can be used for evaluation within CompressAI: JPEG, JPEG2000, WebP, BPG, HEVC, AV1, VVC.

This paper also reports objective comparisons with other implementations and traditional codecs using PSNR and MS-SSIM metrics vs. bitrate, using the Kodak5 [7] image dataset as the test set.

Several extensions to CompressAI are planned. In the next releases, CompressAI will include additional models from the state-of-the-art on learned image compression, and new pre-trained weights for perceptual metrics (e.g.: MS-SSIM). One critical envisioned extension is to add support for video compression. In particular, CompressAI is expected to support the evaluation of traditional video compression standards codecs and end-to-end networks with compressible motion information modules in low-delay and random-access configurations.
The platform is made available to the research and open source communities, under the Apache license version 2.0. We plan to continue supporting and extending CompressAI publicly on GitHub, and we welcome feedback, questions, and contributions.

*********************************************************

## Simplified entropy model for reduced-complexity end-to-end variational auto-encoder with application to on-board satellite image compression

Mr. Vinicius Alves De Oliveira[1,2], Mr. Thomas Oberlin[3], Ms. Marie Chabert[1], Mr. Charly Poulliat[1], Mr. Mickael Bruno[4], Mr. Christophe Latry[4], Mr. Mikael Carlavan[5], Mr. Simon Henrot[5], Mr. Frederic Falzon[5], Mr. Roberto Camarero[6]

*[1]University of Toulouse IRIT / INP-ENSEEIHT, Toulouse, France, [2]TéSA, Toulouse, France, [3]University of Toulouse ISAE-SUPAERO, Toulouse, France, [4]CNES, Toulouse, France, [5]Thales Alenia Space, Cannes, France, [6]ESA-ESTEC, Noordwijk, Netherlands*

In recent years, neural networks have emerged as data-driven tools to solve problems which were previously addressed with model-based methods. In particular, image processing has been largely impacted by convolutional neural networks (CNNs). Recently, CNN-based auto-encoders have been successfully employed for lossy image compression [1,2,3,4]. These end-to-end optimized architectures are able to dramatically outperform traditional compression schemes in terms of rate-distortion trade-off. The auto-encoder is composed of an encoder and a decoder both learned from the data. The encoder is applied to the input data to produce a latent representation with minimum entropy after quantization. The latent representation, derived through several convolutional layers composed of filters and activation functions, is multi-channel (the output of a particular filter is called a channel or a feature) and non-linear. The representation is then quantized to produce a discrete-valued vector. A standard entropy coding method uses the entropy model inferred from the representation to losslessly compress this discrete-valued vector. A key element of these frameworks is the entropy model. In earlier works [1,2,3], the learned representation was assumed independent and identically distributed within each channel and the channels were assumed independent of each other, resulting in a fully-factorized entropy model. Moreover, a fixed entropy model was learned once, from the training set, preventing any adaptation to the input image during the operational phase. The variational auto-encoder proposed in [4] proposed to use a hyperprior auxiliary network. This network estimates the hyper-parameters of the representation distribution, for each input image. Thus, it does not require the assumption of a fully-factorized model which conflicts with the need for context modeling. This variational auto-encoder achieves compression performance close to the one of BPG (Better Portable Graphics) at the expense of a considerable increase in complexity.

However, in the context of on-board compression, a trade-off between compression performance and complexity has to be considered to take into account the strong computational constraints. For this reason, the CCSDS (Consultative Committee for Space Data Systems) lossy compression standard has been designed as a highly simplified version of JPEG2000. This work follows the same logic, however in the context of learned image compression. The aim of this paper is to design a simplified version of the variational auto-encoder proposed in [4] in order to meet the on-board constraints in terms of complexity while preserving high performance in terms of rate-distortion. Apart from straightforward simplifications of the transform (e.g. reduction of the number of filters in the convolutional layers), we mainly propose a simplified entropy model that preserves the adaptability to the input image.

A preliminary reduction of the number of filters reduces the complexity by 62% in terms of FLOPs with respect to [4]. It also reduces the number of learned parameters with a positive impact on the memory occupancy. The entropy model simplification exploits a statistical analysis of the learned representation for satellite images, also performed in [5] for natural images. This analysis reveals that most of the features are well fitted by centered Laplacian distributions. The complex hyperprior model based on a non-parametric distribution of [4] can thus be replaced by a simpler parametric centered Laplacian model. The problem then amounts to a classical and simple estimation of a single parameter referred to as the scale. Our simplified entropy models reduces the complexity of the variational auto-encoder coding part by 22% and outperforms the end-to-end model proposed in [1] for the high target rates.

Acknowledgment: This work has been carried out under the financial support of the French space agency CNES and Thales Alenia Space.

References

[1] Ballé, Laparra, Simoncelli, “End-to-end optimized image compression,” ICLR 2017.
[2] Rippel, Bourdev, “Real-time adaptive image compression,” ICML 2017.
[3] Theis, Shi, Cunningham, Huszar, “Lossy image compression with compressive autoencoders,” ICLR 2017.
[4] Ballé, Minnen, Singh, Hwang, Johnston, “Variational image compression with a scale hyperprior,” ICLR 2018.
[5] Dumas, Roumy, Guillemot, “Autoencoder based image compression: Can the learning be quantization independent?” ICASSP 2018.