Scalable Machine Learning with Heat: Enhancing Large-Scale Anomaly Detections

Hakan Akdag, Fabian Hoppe, Wadim Koslow, Kathrin Rack, and Alexander Rüttgers
German Aerospace Center (DLR), Institute of Software Technology
Cologne, Germany

The *RESIKOAST* project at DLR

Need for resilient coastal protection

- Due to **climate change**, coastal regions are facing **increased threats** like rising sea levels and extreme weather events
- Thus, **enhancing the resilience** of these regions is critical to **protect ecosystems and communities** from long-term adverse impacts

Scope of the *RESIKOAST* project at DLR

- The *RESIKOAST* project aims at developing strategies for the long-term adaptation and tools for **early risk detection** and **timely interventions** in order to protect coastal landscapes, populations, and infrastructure at the North and Baltic Sea coasts
- For this early risk detection, the **identification of hotspots**, i.e., locations with **significant changes [1]** or **anomalies** over time, is crucial
- Hotspot detection is for instance based on the application of the well-known **local outlier factor (LOF)** algorithm to time series of SAR images, with a PyTorch-based prototype evaluated on (spatial) subsets of the entire data set

The challenge

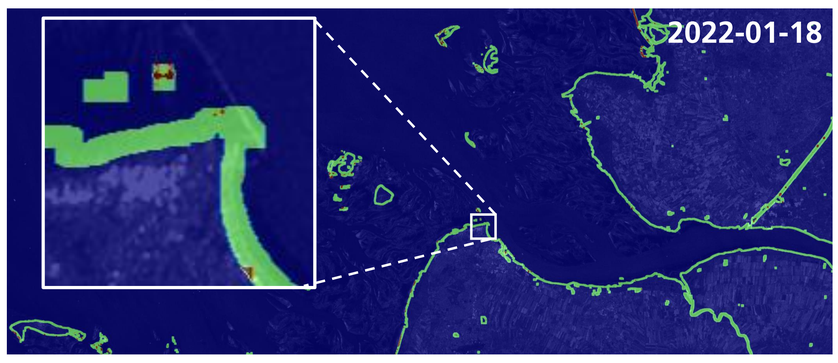
- Size of the currently available data: **~10 TB** already for the North Sea area only (images every 6 days from 2016 to 2023)
- **Scaling up** the prototype to the full (and growing) data set is challenging as on a single workstation/cluster-node…
  …**RAM is exceeded** when processing the entire data set…
  …splitting the data set into smaller parts (that fit into RAM) does not make sense and/or results in **too long overall run time for serial processing** of all parts

HPC as a solution

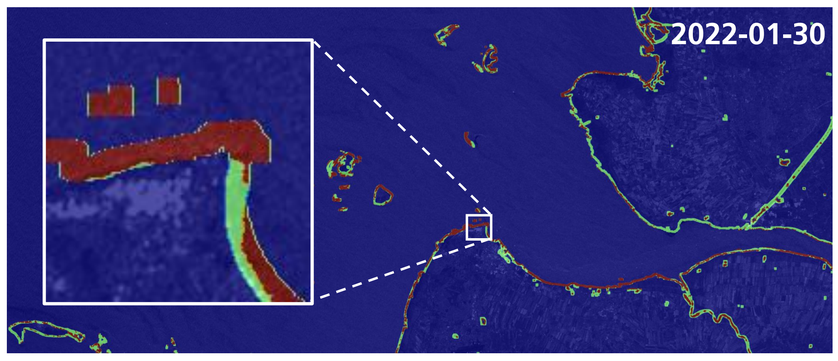
- Tackle the massive amount of data using **multi-node-parallelization** based on the library ***HEAT***, including a memory-efficient implementation of the LOF

Excerpt of results for hotspot detection at the North Sea

At a **„normal" day** (2022-01-18) almost no significant changes are detected, i.e., most pixels of the coastline are green



During the **storm „Nadia"** (2022-01-30) many pixels are classified as anomalous (=red)



The SAR images have been kindly provided by Paola Rozzoli and Luca Dell'Amore from the DLR Microwaves and Radar Institute, Satellite-SAR-Systems Department

Parallel computing framework: *HEAT* [2, 3]

supported by various colleagues from DLR, Research Center Jülich, and Karlsruhe Institute of Technology

The *HElmholtz Analysis Toolkit (HEAT)* is an open source Python library designed for parallel array computing and large-scale machine learning. It enables scientific data analysis on massive datasets by leveraging multi-node/multi-GPU-parallelization and simplifies real-world applications by providing easy-to-use functions as in the NumPy/SciPy ecosystem.

Multi-node- and GPU-capabilities

Operations can be performed in a multi-node-/multi-GPU-setting (e.g., on several nodes of a GPU-cluster)

Simple API and usage

The simple API mimics NumPy/SciPy/scikit-learn and allows for rapid prototyping or adaptation of existing workflows also by HPC-non-experts

Visit us on GitHub

Platform independence / Interoperability

As Heat is mainly based on PyTorch and MPI, it is interoperable, portable, and supports hardware of different vendors (e.g., GPUs by Nvidia & AMD)

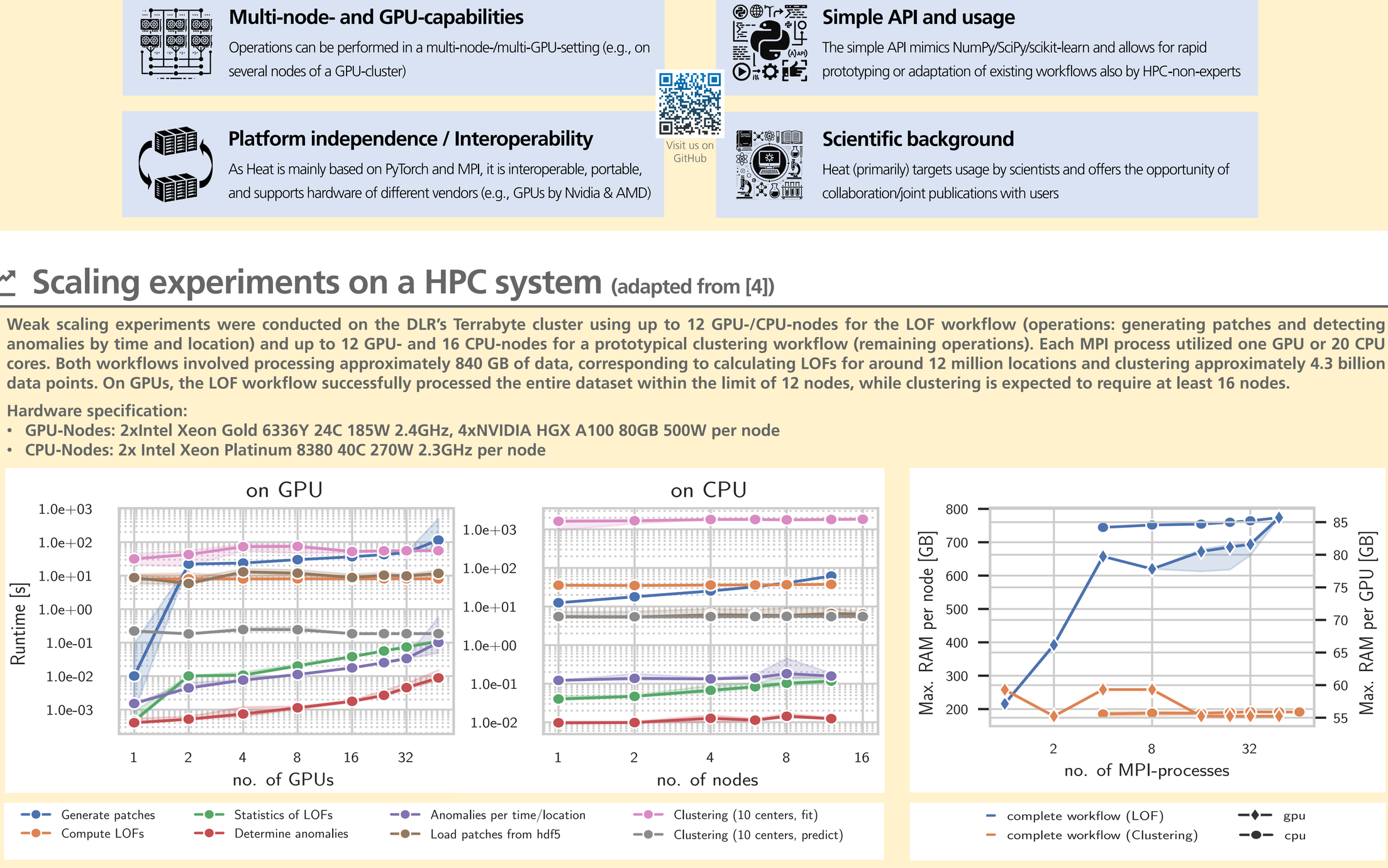
Scientific background

Heat (primarily) targets usage by scientists and offers the opportunity of collaboration/joint publications with users

Scaling experiments on a HPC system (adapted from [4])

Weak scaling experiments were conducted on the DLR's Terrabyte cluster using up to 12 GPU-/CPU-nodes for the LOF workflow (operations: generating patches and detecting anomalies by time and location) and up to 12 GPU- and 16 CPU-nodes for a prototypical clustering workflow (remaining operations). Each MPI process utilized one GPU or 20 CPU cores. Both workflows involved processing approximately 840 GB of data, corresponding to calculating LOFs for around 12 million locations and clustering approximately 4.3 billion data points. On GPUs, the LOF workflow successfully processed the entire dataset within the limit of 12 nodes, while clustering is expected to require at least 16 nodes.

Hardware specification:
- **GPU-Nodes: 2xIntel Xeon Gold 6336Y 24C 185W 2.4GHz, 4xNVIDIA HGX A100 80GB 500W per node**
- **CPU-Nodes: 2x Intel Xeon Platinum 8380 40C 270W 2.3GHz per node**



References

1. W. Koslow, K. Rack, A. Rüttgers, L. Dell'Amore, and P. Rizzoli *Artifact detection in SAR images with AI methods*., EUSAR, 2024
2. F. Hoppe, J.P. Gutiérrez Hermosillo Muriedas, M. Tarnawa, P. Knechtges, B. Hagemeier, K. Krajsek, A. Rüttgers, M. Götz, and C. Comito. *Engineering a large-scale data analytics and array computing library for research: Heat*., ECEASST, 2024.
3. M. Götz et al. *HeAT – a Distributed and GPU-accelerated Tensor Framework for Data Analytics*. In 2020 IEEE International Conference on Big Data (Big Data), pages 276–287, 2020.
4. F. Hoppe, W. Koslow, K. Rack, and A. Rüttgers. *Exploring and processing large data sets in earth observation on HPC-systems with Heat*. 75th International Astronautical Congress, Milan, Italy, 14-18 October 2024.

This research was partially supported by the European Space Agency through the Open Space Innovation Platform (https://ideas.esa.int) as a Early Technology Development Agreement and carried out under the Discovery Program ESA Early Technology Development (Research Agreement No. 4000144045/24/NL/GLC/ov). The view expressed in this publication can in no way be taken to reflect the official opinion of the European Space Agency.