SCALABLE WORKFLOWS FOR REMOTE SENSING DATA PROCESSING WITH THE DEEP-EST MODULAR SUPERCOMPUTING ARCHITECTURE

Ernir Erlingsson[1], *Gabriele Cavallaro*[2], *Helmut Neukirchen*[1], *Morris Riedel*[1,2]

[1] School of Engineering and Natural Sciences, University of Iceland, Iceland
[2] Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany

ABSTRACT

The implementation of efficient remote sensing workflows is essential to improve the access to and analysis of the vast amount of sensed data and to provide decision-makers with clear, timely, and useful information. The Dynamical Exascale Entry Platform (DEEP) is an European pre-exascale platform that incorporates heterogeneous High-Performance Computing (HPC) systems, i.e., hardware modules which include specialised accelerators. This paper demonstrates the potential of such diverse modules for the deployment of remote sensing data workflows that include diverse processing tasks. Particular focus is put on pipelines which can use the Network Attached Memory (NAM), which is a novel supercomputer module that allows near processing and/or fast shared storage of big remote sensing datasets.

Index Terms— Remote Sensing, Modular Supercomputing Architecture (MSA), Network Attached Memory (NAM), High-Performance Computing (HPC), hardware accelerators

1. INTRODUCTION

The continuous developments of remote sensing platforms and sensor technologies combined with the open and free data policy of Earth Observation (EO) programs is generating an unprecedented volume and variety of raw data [1]. Copernicus, with its fleet of Sentinel satellites, is the world's largest single EO programme. For example, the two twin satellites Sentinel 2A and 2B deliver 23 TB/day of multispectral data. Whilst previously the major issue for researchers has been the identification of accessible remote sensing data sources, nowadays the main issue is how to make the processing of this vast abundance of open data scalable [2]. In 2017, the Sentinel Data Access System experienced a publication rate of 10.04 TB/day with an average download volume of 93.5 TB/day[1]. Due to the insufficient memory size and number of cores available in commodity computers on the one hand and on the other hand the increasing number of applications that require data computing in near real time (i.e., supporting decision-makers), remote sensing data processing pipelines necessitate the use of parallel algorithms that can run and scale on High-Performance Computing (HPC) systems with distributed memory [3].

Lately, several distributed architectures have been developed to make HPC computing available for remote sensing including cloud-based systems [4], Message Passing Interface (MPI) systems [5], and Map-Reduce systems [6]. In this context, high-end HPC clusters, that currently reach performance in the order of petaflops (i.e., 10^{15} floating point operations per second), are already delivering unprecedented breakthroughs [7]. However, emerging machine learning and deep learning algorithms are transforming the workloads that are run on HPC systems, with the need for higher memory, storage, and networking capabilities, as well as optimized software and libraries to deliver the required performance. Thus, tomorrow's HPC has to provide heterogeneous hardware accelerators and software technologies within the same architecture in order to cover both the needs of classic HPC simulations and of novel data analytics applications.

While Erlingsson *et al.* [8] already described a strategy for mapping the different phases of a Support Vector Machine (SVM) classifier to the most suitable accelerator modules of the Modular Supercomputer Architecture (MSA), the role of the Network Attached Memory (NAM) module[2], which is one key innovative element of the MSA, has so far only been partially investigated.

This paper focuses therefore in particular on the NAM and its benefits within more general remote sensing data processing workflows. The NAM is a high-speed, non-volatile, network-accessible storage with built-in processing capabilities that is designed for storage of calculations, data, and other operations in progress (e.g., scratchpad or checkpoint-restart space). We describe an experiment for writing and reading tasks using two NAM hardware prototypes. The results demonstrate the efficiency of this module and its potential for performing processing and fast shared storage opera-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 754304 DEEP-EST.

[1] `https://sentinel.esa.int/web/sentinel/news/-/article/sentinel-data-access-annual-report-2017`

[2] `https://www.deep-projects.eu/hardware/memory-hierarchies/49-nam`

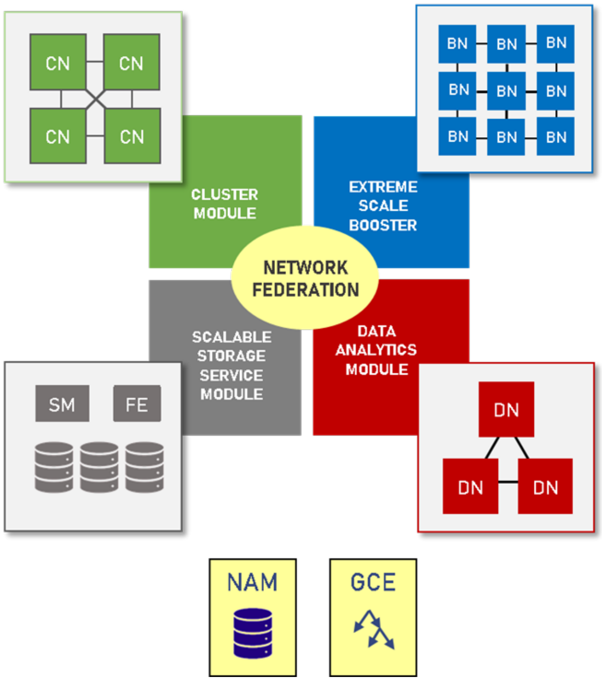

Fig. 1. The Modular Supercomputer Architecture (MSA) under development in the DEEP-EST research project.

tions on data within a certain size limit (see Section 3). Based on this preliminary results, it can be expected that the forthcoming version of the NAM will be able to cope with operations performed on much larger remote sensing data.

2. MODULAR SUPERCOMPUTING ARCHITECTURE

2.1. Dynamical Exascale Entry Platform

The Dynamical Exascale Entry Platform – Extreme Scale Technologies (DEEP-EST) project[3] is the third pre-exascale research project in its series, following the Dynamical Exascale Entry Platforms DEEP and DEEP-ER, being funded by the European Union. The Modular Supercomputer Architecture (MSA) [9] is an innovative HPC architecture that has been developed in the project. It integrates an arbitrary number of modules of heterogeneous hardware components, in particular specialised hardware accelerators. The aggregation of these modules within the same architecture forms the blueprint for future exascale and energy efficient computing systems. Each module of the MSA is custom-built to meet the requirements of a specific set of computation, storage, or communication tasks within a processing pipeline. This includes remote sensing workflows.

2.2. Co-Design of Hardware for Remote Sensing Data

A typical remote sensing workflow starts with data acquisition from a set of sensors and ends with the provision of valuable information to a given thematic application. Such a workflow is usually both data and compute intensive since it is not only challenged by the increasing volume and variety of the data but also by the high computational complexity of many data mining algorithms. When considering preprocessing and information extraction tasks (e.g., clustering, classification, etc.) that adopt either classical machine learning algorithms or more advanced deep learning methods, one of the main challenge is to make them exploit parallel computing systems efficiently. In the context of the DEEP-EST project, this is tackled by a co-design approach, i.e. the MSA hardware and system software are designed to match the requirements of the targeted scientific applications. The expected outcome is an efficient assignment of heterogeneous processing tasks to the most suitable modules of the MSA. The MSA modules (cf. Fig. 1) support the different parts of a remote sensing workflow as follows:

The Cluster Module (CM) includes the fastest Central Processing Units (CPUs), which makes it suitable for tasks that are the most computationally expensive but with limited scalability (e.g., classification algorithms with elevated complexity that can be performed with high single-thread performance).

On the other hand, the Extreme Scale Booster (ESB) [9] module can be described as putting emphasis on scaling parallel tasks. It consists of many powerful Graphics Processing Unit (GPU) accelerators where a GPU is coupled to a rather weak host CPUs (i.e. low performance and lower amounts of memory) in a node: these CPUs are only needed for offloading communication with other ESB nodes and/or enabling I/O. In terms of memory, the ESB will take advantage mostly of the fast GPU RAM and will perform communication via GPUDirect to other accelerators to avoid the bottleneck of the host CPU memory. The performance scalability that can be achieved in the DEEP-EST ESB is expected to be higher than the one thatcan be reached using standard technologies such as NVlink or NVSwitch[4]. In fact, the process of scaling is not restricted to the GPUs within a node, but occurs also across nodes (i.e., availability of a higher number of accelerators that are efficiently interconnected). This is an important factor when deep learning algorithms are applied in remote sensing, since large size datasets (e.g., Sentinel 2 data) require scaling beyond the single node especially when large-scale inference operations are performed.

The ESB also integrates in its fast interconnection network federation fabric the Global Collective Engine (GCE) that is an accelerator for speeding-up Message Passing Interface (MPI) collective operations in hardware, e.g., summing

[3] https://www.deep-est.eu/

[4] https://www.nvidia.com/en-us/data-center/nvlink/

up values transmitted in MPI messages.

While the Data Analytics Module (DAM) consists of many GPU accelerators just like the ESB module, it differs from the ESB in having more powerful CPUs, more RAM, fast local non-volatile memory, and allowing extra-acceleration via Field-Programmable Gate arrays (FPGAs). The DAM serves the purpose of enabling data-intensive computing that requires large memory capacity (e.g., store huge amounts of weights and activations of deep networks). The DAM enables also more complex workflows that can benefit in addition to the GPUs from the more powerful CPUs than those available in the ESB nodes.

The Network Attached Memory (NAM) and the Scalable Storage Service Module (SSSM) are two further modules that are included in the MSA. They are described in more details in the next section.

3. NETWORK ATTACHED MEMORY MODULE

3.1. NAM Introduction

Performance slowdown due to frequent and/or slow storage access is a known problem in contemporary super-computing. Big data sets are usually stored in a centralized storage module, such as the Scalable Storage Service Module (SSSM) This module is usually located in a separate chassis, physically separate from the computing nodes which –despite fast interconnects– slows down data access speed due to latency. To mitigate this problem it is common practice to include additional storage directly in each chassis, such as a local Solid State Drives (SSDs). However, this is not sufficient when dealing with truly big data, or data that has to be continuously shared among different modules and/or nodes.

The NAM module [10] addresses this problem by serving as an extremely fast storage target using a Hybrid Memory Cube (HMC) that is directly fused with the fabric itself, close to the computational nodes. (HMC is a new class of non-volatile memory, approx. 1000 times faster than SSDs.) Furthermore, the NAM is equipped with an FPGA which can be used as an accelerator for fast near-data processing.

3.2. NAM: Preliminary Results

To investigate the benefits of using the NAM module within a processing workflow, experiments were conducted on limited hardware prototypes since at the time of writing, the final MSA is still being built, including a future, improved NAM.

An evaluation of two of these NAM prototypes was performed by using them as storage targets of 4 GB capacity, with a maximum capacity of 2 GB available for each prototype. Figure 2 depicts the read and write throughput of the NAM compared to the BeeGFS parallel file system[5] that accesses a remote file server. In both cases, the experiments were performed using the same nodes of the DEEP-ER supercomputer. It should be noted that it was not possible to fully bypass the use of caching for BeeGFS read operations as only the local operating system honored the request to disable caching, but not necessarily other parts of the BeeGFS chain. Therefore, since the displayed throughput is the result of multiple read operations that are averaged using the same set of files, the BeeGFS read measurements are in fact far too optimistic and the true read performance can be considered to be slower by an unknown extent. However, this not the case for the file system write operations, which were successfully measured while completely bypassing the use of caching.

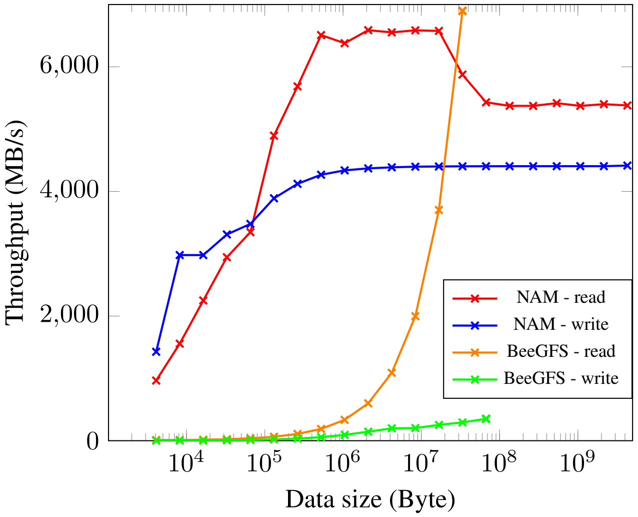

Fig. 2. I/O throughput comparison between the NAM and the BeeGFS parallel file system.

As Figure 2 shows, the NAM throughput is significantly better, in particular with respect to write operations, but even still outperforming the file system with caching when reading files up to 16 MB in size. The NAM performance slump during read operations is due to an internal buffer of the same size: after this buffer has been filled, further buffers must be allocated, which negatively affects the overall read performance. However, more allocations do not slow down the performance any further.

These results show the potential of using the NAM as a storage target rather than the standard file system. For the given prototype, this applies especially for data smaller than 32 MB, which is, e.g., suitable for computer vision applications using batches of multi-spectral remote sensing data to train deep learning networks. For the forthcoming DEEP-EST NAM, however, the results are expected to vastly improve once hardware and software problems encountered with the prototype are resolved. It is expected that a production ready NAM will produce a throughput of over 10 GB a second as it will, among other improvements, double the number of interconnect links. Furthermore, its HMC storage capacity is extended to 128 GB, with an additional (but slower) flash

[5] https://www.beegfs.io/content/

memory that offers storage in the Terabyte range.

The potential of the NAM was also confirmed by evaluating the prototype with checkpointing code routines from the libNAM library[6], which produced results that were consistent to the graph in Fig. 2. The usage of checkpointing increases application robustness by allowing them to resume from a certain execution snapshot (checkpoint) instead of having to start over if the application terminates unexpectedly. For checkpointing, the NAM's FPGA is also utilized to perform parity checkpointing on the NAM rather than on a compute node.

Preliminary results indicate that workflows that process big remote sensing data will benefit from the NAM, especially when considering that remote sensing data are often multi-modal, e.g. from optical (multi- and hyperspectral) and synthetic aperture radar (SAR), which require ensembles of machine learning and/or deep learning models to combine and fuse them. The forthcoming NAM, with increased throughput, local computational ability, and massive storage capability, will be able to process several intermediate multi-modal learning models simultaneously and boost the whole processing pipeline.

4. CONCLUSIONS

This paper described the potential of the accelerators of the Modular Supercomputer Architecture (MSA) that can be tailored to specific applications, such as for realising scalable remote sensing data processing workflows. Since the MSA is under development, only preliminary results were provided for using the Network Attached Memory (NAM) module as a very fast remote storage that can be shared between nodes of an HPC cluster, e.g. for passing data from one stage of a workflow to the next stage. The throughput of the NAM prototype is impressive and was only topped by a remote file system due to caching of previously read data. The final NAM version is expected to perform even better. It will have a larger storage capacity and thus allows remote sensing workflows to store, e.g. intermediate machine learning model data in order to speed up, e.g., scalable SVMs on the MSA [8].

5. REFERENCES

[1] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big Data for Remote Sensing: Challenges and Opportunities," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016.

[2] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie, "Remote Sensing Big Data Computing: Challenges and Opportunities," *Future Generation Computer Systems*, vol. 51, pp. 47–60, 2015.

[3] A. Plaza, Q. Du, Y. Chang, and R. L. King, "High Performance Computing for Hyperspectral Remote Sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 3, pp. 528–544, 2011.

[4] S. You, J. Zhang, and L. Gruenwald, "Large-Scale Spatial Join Query Processing in Cloud," in *31st IEEE International Conference on Data Engineering Workshops, ICDE Workshops 2015*. 2015, IEEE.

[5] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson, and A. Plaza, "On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4634–4646, 2015.

[6] V. A. Ayma, R. S. Ferreira, P. Happ, D. Oliveira, R. Feitosa, G. Costa, A. Plaza, and P. Gamba, "Classification Algorithms for Big Data Analysis, a Map Reduce Approach," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 2015.

[7] R. McKinney, V. K. Pallipuram, R. Vargas, and M. Taufer, "From HPC Performance to Climate Modeling: Transforming Methods for HPC Predictions Into Models of Extreme Climate Conditions," in *Proceedings - 11th IEEE International Conference on eScience, eScience*, 2015.

[8] E. Erlingsson, G. Cavallaro, M. Riedel, and H. Neukirchen, "Scaling Support Vector Machines Towards Exascale Computing for Classification of Large-Scale High-Resolution Remote Sensing Images," in *2018 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018*, 2018, pp. 1792–1795.

[9] N. Eicker, "Taming Heterogeneity in HPC," Keynote Presentation at SAI Computing Conference 2016, July 2016, https://youtu.be/aM9AkgG5ud4.

[10] J. Schmidt, "NAM – Network Attached Memory," Doctoral Showcase poster at International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2016, Nov. 2016, http://www.deep-projects.eu/images/nam_poster_SC16.pdf.

[6] https://gitlab.version.fz-juelich.de/galonska1/libNAM