PROCEEDINGS OF SCIENCE

Baler: Machine Learning-based Data Compression

Axel Gallén[a,*] **and James Smith**[b] **for the Baler collaboration**

[a]*Department of Physics and Astronomy, Uppsala University,*
Läderhyggsvägen 1, Uppsala, 752 37, Sweden

[b]*Department of Physics and Astronomy, The University of Manchester,*
Schuster Building, Oxford Road, Manchester, M13 9PL, United Kingdom

E-mail: axel.lars.gallen@cern.ch, james.smith-7@manchester.ac.uk, baler-compression-members@cern.ch

With the rise of novel computing techniques such as big data and artificial intelligence, many scientific and industrial disciplines are faced with exponentially increasing demands for data storage and compute resources. Traditional data compression algorithms are either generically applicable but lossless, limiting performance (e.g. `zip`), or lossy but designed for specific applications (e.g. `jpg`, `mp3`). Machine learning can be deployed to identify the most significant features of any given dataset and favour these features in a compression algorithm. Baler [1] is a novel framework for developing, testing and deploying autoencoder-based data compression algorithms. In this talk, we report on recent developments to the Baler framework, including the implementation of Baler on FPGAs and how Baler has been used to compress data from atomic physics (Mössbauer imaging).

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

[*] Speaker

1. Introduction

Many scientific disciplines face a common challenge: managing increasingly large datasets. For example, by the end of 2030, the Large Hadron Collider (LHC) and its experiments are projected to analyze more than ten times what is currently being handled [2–4], while fields like Computational Fluid Dynamics (CFD) and atomic physics (x-ray imaging) rely on simulation and imaging methods which produces terabytes of data, all in need of storage and distribution. Without significant R&D efforts, the vast amount of data generated by these large-scale scientific projects is expected to surpass the available storage capacities (see e.g. Fig 1 regarding the ATLAS experiment at the LHC). This issue is noonly confined to scientific research but affects industrial sectors as well [5].

In addition to storage, the efficient distribution of big data is another critical concern. Data compression techniques can play a key role in reducing the volume of data that needs to be transmitted, making distribution more manageable – and faster. Hardware solutions like Field-Programmable Gate Arrays (FPGAs) are increasingly being explored for real-time data compression due to their high processing speed and flexibility. This can significantly optimize both storage and data sharing processes in large-scale scientific and industrial applications.

The most common mitigation strategy to this problem involves so-called *lossless* data compression. With this strategy, one is able to compress data down to a size capped by the compression algorithm, then forcing discarding of data past that point. For a more aggressive approach, one can delve into *lossy* data compression. This method uses approximations and the partial discarding of data to reduce data size further than lossless methods at the cost of data fidelity. Performing lossy compression using Autoencoders (AEs) [6], has been proven to be viable in many different fields [7] including High Energy Physics (HEP) [8–10], and a representation of the Autoencoder neural network model is presented in Fig 2. This is the core part of the Baler framework, which has been developed to easily investigate the feasibility of lossy compression using machine learning methods. For a more detailed description of how Baler incorporates and uses AEs for data compression, as well as for previous results in the CFD and HEP fields, the reader is referred to Reference [11].

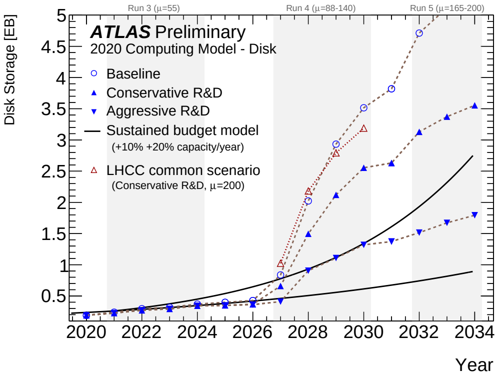

Figure 1: Projected evolution of disk usage from 2020 until 2034 at the ATLAS experiment [12].

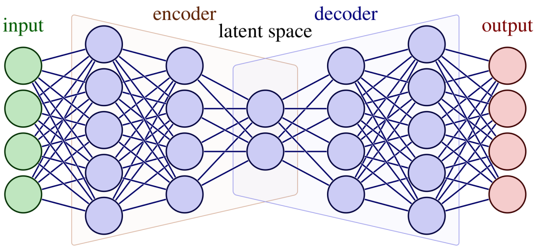

Figure 2: Visualization of an Autoencoder neural network. Modified from [13].

2. Mössbauer

When a photon is emitted from a nuclide undergoing an energy level transition, the energy of the photon is that of the difference in energy levels, less the recoil energy imparted on the nuclide. For free atoms this recoil energy can be significant and can inhibit direct absorption of the photon by another nuclide of the same type. However, when the nuclide is in a lattice the effective mass of the nuclide becomes large and the recoil energy imparted becomes small. At this point, the photon energy is similar to that of the transition energy and resonant absorption and emission becomes possible. This is known as the *Mössbauer effect* [14], and produces a characteristic resonant peak in the absorption spectra.

The initial gamma ray source can be linearly oscillated to account for any remaining recoil energy in a technique known as *Mössbauer spectroscopy*, and can be used to study the hyperfine transition in atoms. A *Mössbauer camera* can be constructed by placing a CMOS sensor or similar behind the sample to be measured in the spectroscope. By using the CMOS output to tune the oscillation of the gamma-ray source, the structure of the material can be directly imaged via the Mössbauer effect, allowing for a resolution as small as 5 µm. [15, 16]

One such camera with a 2048x2048 pixel sensor would produce a data rate of 2-4 MB/s during a typical data-taking period of 1-5 days. This produces up to 2 TB of data per period. The data can be compressed down to 10% of the original size with an acceptable loss of quality. Background noise is significantly reduced whilst the features of the image are sufficiently preserved. An example can be seen below in Figure 3.

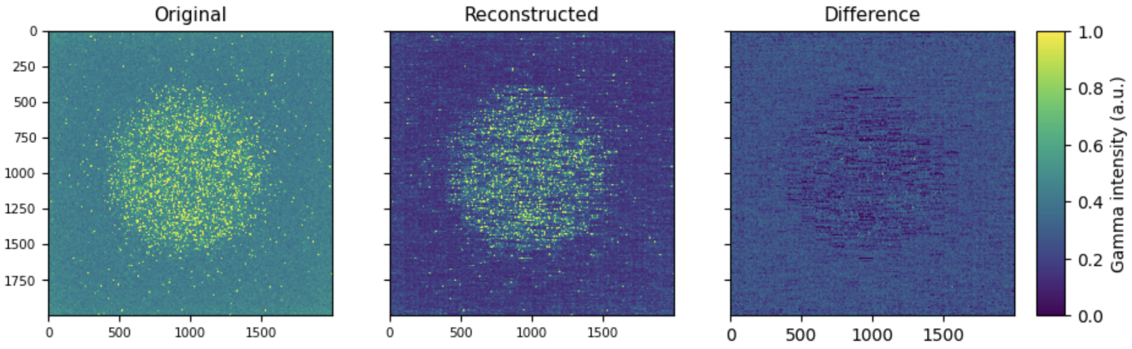

Figure 3: A comparison of an image captured by a Mössbauer camera before and after compression. The features of the image (dots) are preserved at a good quality, and the background noise is reduced.

3. FPGA Compression

Implementing machine-learning-based data compression on FPGAs provides a high-performance, efficient solution that leverages hardware acceleration. With inherent parallelism and low-latency processing, FPGAs are ideal for the computational demands of ML algorithms, particularly in real-time compression tasks for data-heavy applications like video/data streaming and telecommunications where bandwidth efficiency is crucial. Autoencoders and similar models can be fine-tuned on FPGAs to reduce data size before transmission, thereby optimizing bandwidth usage. FPGAs' parallel processing also accelerates complex compression tasks that are otherwise impractical on traditional processors, making them highly effective for bandwidth compression in resource-constrained environments.

Figure 4 describes the general workflow/flowchart that was followed when investigating machine-learning-based data compression on FPGAs. One of many results obtained by A. Lilius [17] is presented in Table 1. These results show the throughput of three different DNN models on an FPGA compared to a CPU. For this study, it was shown that a throughput increase of up to 16.9x could be obtained by using an FPGA.

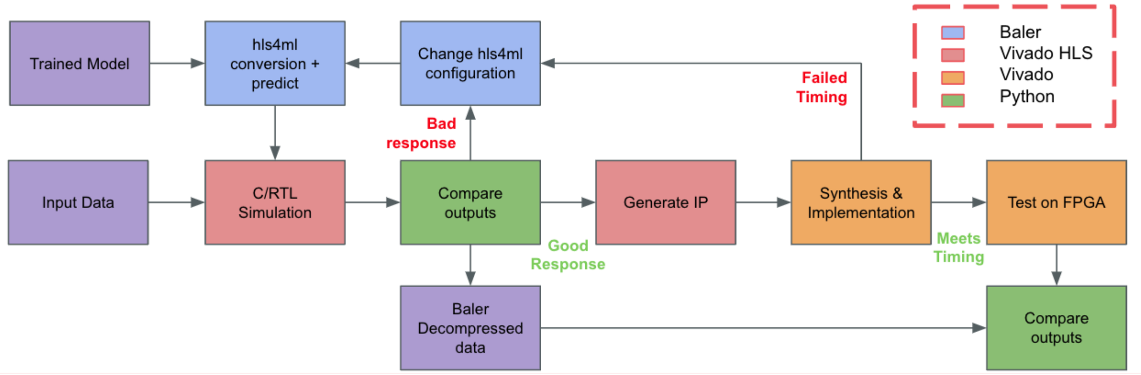

Figure 4: Overview of the workflow which was used to investigate the viability of machine-learning-based data compression on FPGAs.

Model (Encoder)	Processing Unit	Time (s)	Throughput (inferences / s)
DNN Large	CPU	0.95	189473
	FPGA	1.26	142377
DNN Reduced	CPU	0.94	191489
	FPGA	0.23	768481
DNN Tiny	CPU	0.86	209302
	FPGA	0.05	3472422

Table 1: The results of comparing the implementation of different model sizes for the encoder. From [17]

4. Conclusions & Outlook

In conclusion, managing the growing data demands across scientific and industrial fields requires innovative approaches to storage and distribution. Large-scale projects, like those at the LHC, produce vast amounts of data that current infrastructure struggles to handle. Lossy data compression methods offer promising solutions by reducing data size - e.g. by using Autoencoders. Moreover, leveraging FPGAs for machine-learning-based real-time data compression presents a high-performance, efficient solution. An FPGA's parallel processing capabilities significantly optimize compression tasks, improving throughput and bandwidth efficiency for data-intensive applications.

In the future, the Baler collaboration will continue to advance and refine techniques for machine-learning-based data compression, with a particular focus on optimizing performance for large-scale scientific and industrial applications. A key area of interest will be the integration of FPGAs, which have shown significant promise and potential for real-time data compression.

References

[1] Baler collaboration, *Baler: a machine learning based data compression tool*, version 1.4.0, Apr. 2023. DOI: 10.5281/zenodo.10723669. Available at: https://github.com/baler-collaboration/baler.

[2] ATLAS Collaboration, *ATLAS Software and Computing HL-LHC Roadmap*, CERN-LHCC-2022-005, LHCC-G-182, CERN, Geneva, 2022. http://cds.cern.ch/record/2802918.

[3] CMS Offline Software and Computing, *CMS Phase-2 Computing Model: Update Document*, CMS-NOTE-2022-008, CERN-CMS-NOTE-2022-008, CERN, Geneva, 2022. http://cds.cern.ch/record/2815292.

[4] LHCb Collaboration, *Computing Model of the Upgrade LHCb experiment*, CERN-LHCC-2018-014, LHCB-TDR-018, CERN, Geneva, 2018. http://cds.cern.ch/record/2319756, doi: 10.17181/CERN.QOP4.57ON.

[5] Maqbool Khan, Xiaotong Wu, Xiaolong Xu, and Wanchun Dou, *Big data challenges and opportunities in the hype of Industry 4.0*, in Proceedings of the 2017 IEEE International Conference on Communications (ICC), 2017, pp. 1-6. doi: 10.1109/ICC.2017.7996801.

[6] Mark A. Kramer, *Nonlinear principal component analysis using autoassociative neural networks*, AIChE Journal, vol. 37, no. 2, pp. 233-243, 1991. doi: 10.1002/aic.690370209. https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690370209.

[7] Tong Liu, Jinzhen Wang, Qing Liu, Shakeel Alibhai, Tao Lu, and Xubin He, *High-Ratio Lossy Compression: Exploring the Autoencoder to Compress Scientific Data*, IEEE Transactions on Big Data, vol. 9, no. 1, pp. 22-36, 2023. doi: 10.1109/TBDATA.2021.3066151.

[8] Sten Åstrand, *Autoencoder Compression in High Energy Physics*, Student Paper, 2022. http://lup.lub.lu.se/student-papers/record/9075881

[9] Axel Gallén, *An Open-Source Autoencoder Compression Tool for High Energy Physics*, Student Paper, 2023. https://lup.lub.lu.se/student-papers/record/9117991

[10] Fritjof Bengtsson, Caterina Doglioni, Per Alexander Ekman, Axel Gallén, Pratik Jawahar, Alma Orucevic-Alagic, Marta Camps Santasmasas, Nicola Skidmore, and Oliver Woolland, *Baler – Machine Learning Based Compression of Scientific Data*, 2023. arXiv:2305.02283.

[11] Fritjof Bengtsson Folkesson, Caterina Doglioni, Per Alexander Ekman, Axel Gallén, Pratik Jawahar, Marta Camps Santasmasas, and Nicola Skidmore, *Baler - Machine Learning Based Compression of Scientific Data*, EPJ Web of Conferences, vol. 295, p. 09023, 2024. doi: 10.1051/epjconf/202429509023. https://doi.org/10.1051/epjconf/202429509023.

[12] ATLAS Collaboration, *ATLAS HL-LHC Computing Conceptual Design Report*, CERN-LHCC-2020-015, LHCC-G-178, CERN, Geneva, 2020. https://cds.cern.ch/record/2729668.

[13] Izaak Neutelings, *Neural Networks*, 2021. https://tikz.net/neural_networks/.

[14] R. L. Mössbauer *Kernresonanzfluoreszenz von Gammastrahlung in Ir*^{191} *Z. Physik* **151**, 124–143 (1958) https://doi.org/10.1007/BF01344210

[15] Yoshida, Y., Hayakawa, K., Yukihira, K. et al. *Development and applications of "Mössbauer cameras"*. Hyperfine Interact 198, 23–29 (2010). https://doi.org/10.1007/s10751-010-0228-x

[16] Kobayashi, M., Hayakawa, K., Yoshida, Y. et al. *Mössbauer spectroscopic camera for operando measurement of pattern formation processes.* Hyperfine Interact 243, 9 (2022). https://doi.org/10.1007/s10751-022-01796-0

[17] Aleko Lilius, *Performance of ML-Based Bandwidth Compression on FPGAs*, Student Paper, 2024. http://lup.lub.lu.se/student-papers/record/9169526

A. Baler Collaboration

Khwaish Anjum[4,5]
Fritjof Bengtsson[6]
Marta Camps Santasmasas[7]
Leonid Didukh[6]
Caterina Doglioni[1]
Malena Duroux[1]
Alexander Ekman[6]
Jacob Forsell[6]
Axel Gallén[2]
Elena Gramellini[1]
Samuel Hill[1]
Pratik Jawahar[1]
Yuyang Jin[6]
Kaarel Kvisalu[1]
Aleko Lilius[6]
Nicole Skidmore[3]
James Smith[1]
Chakravarty Varadarajan[1]

1: University of Manchester 2: Uppsala University 3: University of Warwick 4: FSU Jena 5: GSI 6: No current affiliation 7: University of Salford