

Sustainable AI Data Compression: Profiling BOA on BEDE

Measuring the energy and carbon footprint of BOA Constrictor, a neural lossless compressor for particle-physics data

Hanzila Hussain

Supervised by: Caterina Doglioni



The University of Manchester

Background

- **The growing nature of data:**
 - High Energy Physics (HEP) experiments such as the LHC record several petabytes of data every week
 - Storing it comes at a significant cost, in both budget and environmental resources
 - Environmental cost: **electricity** to process data, **embodied carbon** to produce disk server
- **BOA Constrictor: ML-based data compression**
 - developed by researchers at the University of Manchester [1]
 - compression can save on carbon costs due to buying fewer disk servers
- **The scientific question: what is the trade-off between data processing and storage?**
 - ML-based compression is generally more computationally intensive than standard algorithms like ZSTD and LZMA
 - This raises the question of where the **break-even point** lies

Project:

- Compare BOA's carbon footprint with classical compressors
- Measure how much the choice of GPU changes the cost
- Ultimately, build a **learning guide** for understanding break-even point questions

Measuring carbon

Carbon Estimation framework:

Total carbon cost = Operational carbon + embodied carbon

We only measure operational term

$$C = E \times I$$

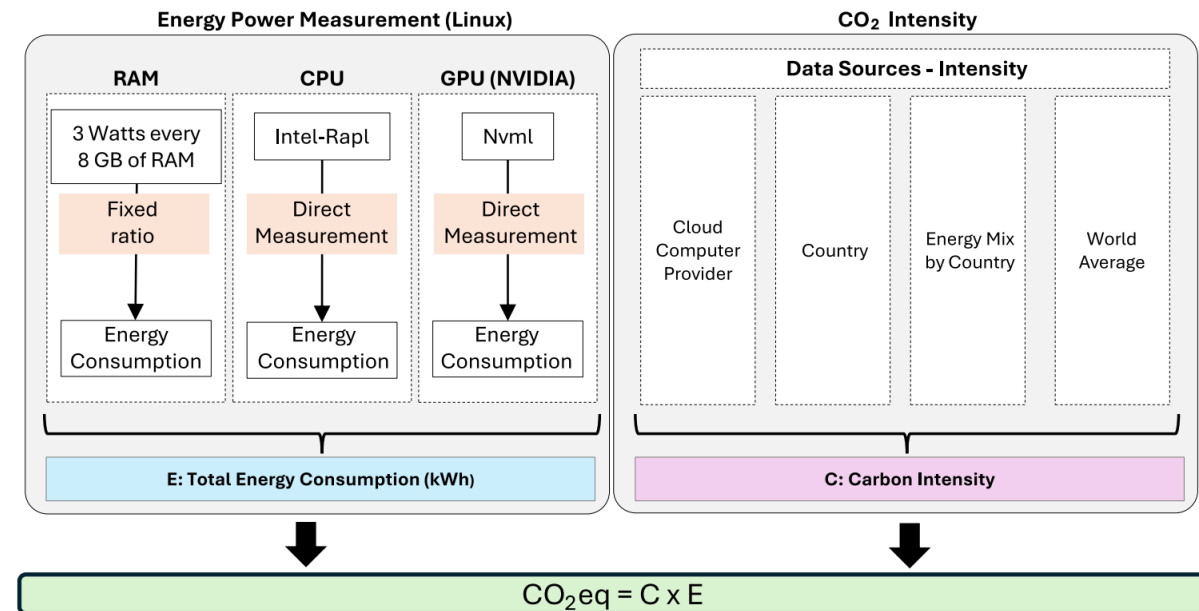
carbon = energy the code used × how dirty the grid is

What CodeCarbon measures

- GPU: board power read from the driver (NVML)
- CPU: RAPL counters, or a TDP table when virtualized
- RAM: estimated from installed capacity

What it assumes

- I: UK grid intensity, looked up offline (GBR)
- Operational energy only - embodied carbon excluded



Measured per stage: Training (per epoch), compression and decompression. Each block sits in one track() context manager, with CUDA synchronised at both ends and power sampled every second.

Data & Method

- **The data** used was a flat float32 array of reconstructed jets from CERN open data.
- BOA model fixed for all experiments
- UK grid intensity was used

Bede (N8CIR)

4 x Tesla V100-SXM2

Workstation

1 x NVIDIA RTX 5090

Google Colab

1 x Tesla T4

What each run records

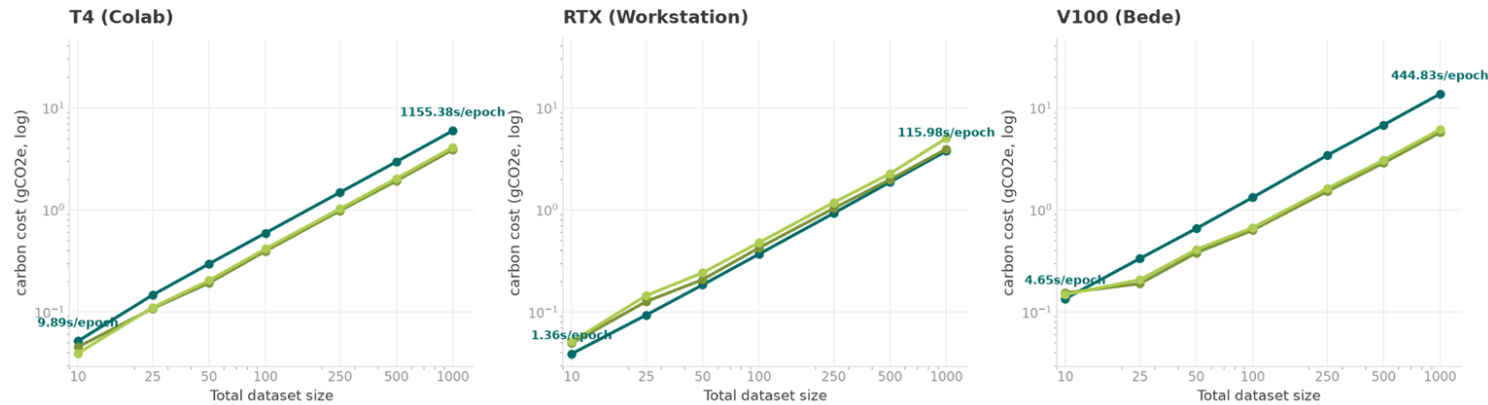
- Seven input sizes per machine: 10, 25, 50, 100, 250, 500 and 1000 MB
- A 30 s idle baseline, then the energy of every training epoch (3 epochs per size)
- Compression and decompression at the final epoch
- One CSV row per stage: GPU/CPU/RAM energy split, kWh per GB, MB/s and a hardware fingerprint

Classical baseline: the same data compressed through zstd and LZMA on a MacBook i5 CPU, into the same CSV schema.

Results

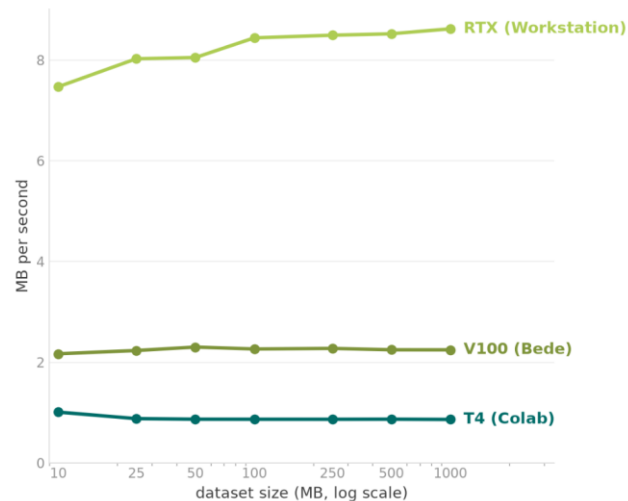
Carbon cost scales with input size across devices

— training (per epoch) — compression (full run) — decompression (full run)

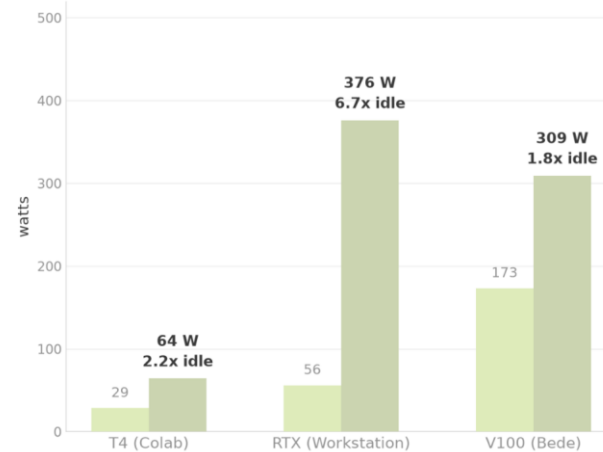


- Carbon cost linear with total size of dataset
- Longer runtime for training doesn't mean more Carbon emission

Training throughput per epoch

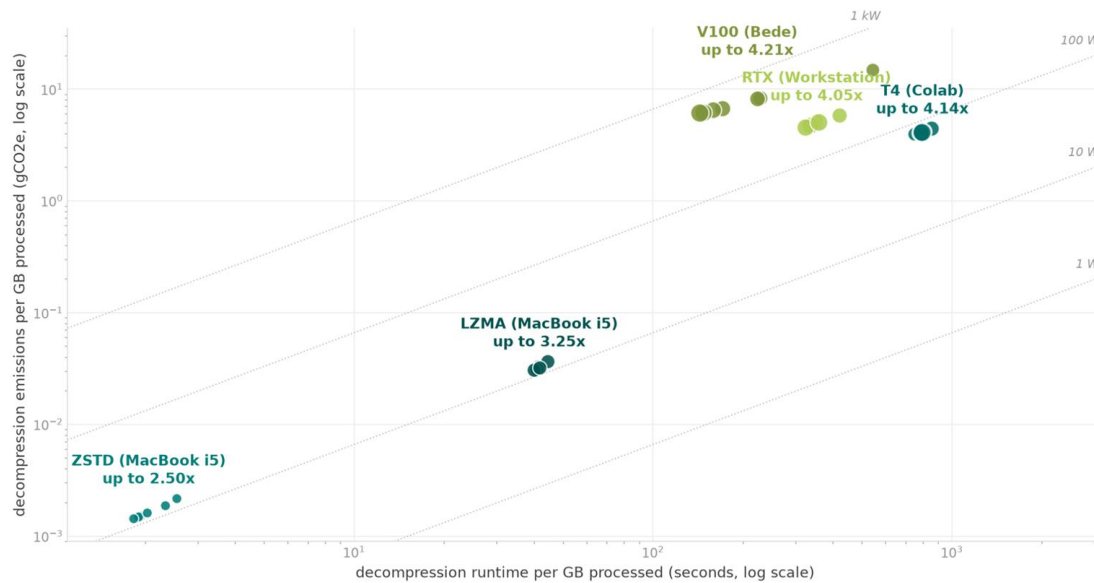


GPU power: idle vs a training epoch

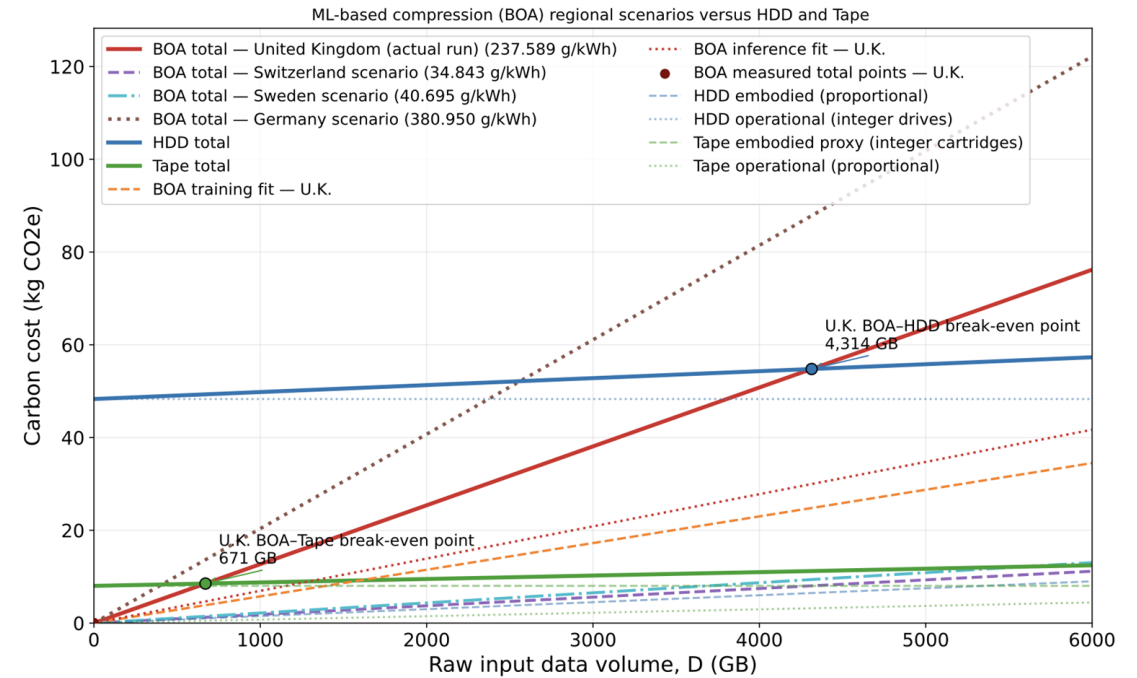


- BOA on V100 devices not parallelised at the time of testing, hence low throughput
→ higher power draw leads to worse carbon impact
- V100 being a cluster is shared

Results



Classical compressors outperform BOA in terms of carbon cost



High compression ratios from BOA can reduce carbon if used at scale

Next steps:

- The team have started mapping a pareto plot for different environments and hyperparameters
- Obtain more realistic estimates from Bede as we need to parallelise the workload better to utilise all devices
- Accelerate inference using Bit Mamba and test carbon cost

Thank you

With thanks to my supervisors **Caterina Doglioni and Akshat Gupta** and to N8CIR for hosting this internship.

References

[1] BOA Constrictor paper

arxiv.org/abs/2511.11337

Code and learning guide

github.com/HanzilaH/N8CIR-SusTrain-data-compression-notebooks