

Lossy Data Compression Exploration in an Online Laboratory and the Link to HPC Design Decisions

Karsten Peters-von Gehlen¹,
Juniper Tyree²,
Sara Faghieh-Naini³,
Peter Dueben³,
Jannek Squar⁴ and
Anna Fuchs⁴

¹ Deutsches Klimarechenzentrum (DKRZ), Hamburg, Germany (peters@dkrz.de); ² University of Helsinki, Institute for Atmospheric and Earth System Research, Helsinki, Finland; ³ European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany; ⁴ Deutsches Klimarechenzentrum (DKRZ), Hamburg, Germany



Backdrop

How voluminous is a state-of-the-art coupled Earth System Model Simulation (5km, 5yrs, 3-hourly output) (using loss-less compression)?

~480TB per simulation

Scientific projects can **easily produce around 10 simulations** and keep them on **warm storage for analysis**, e.g. during hackathons:

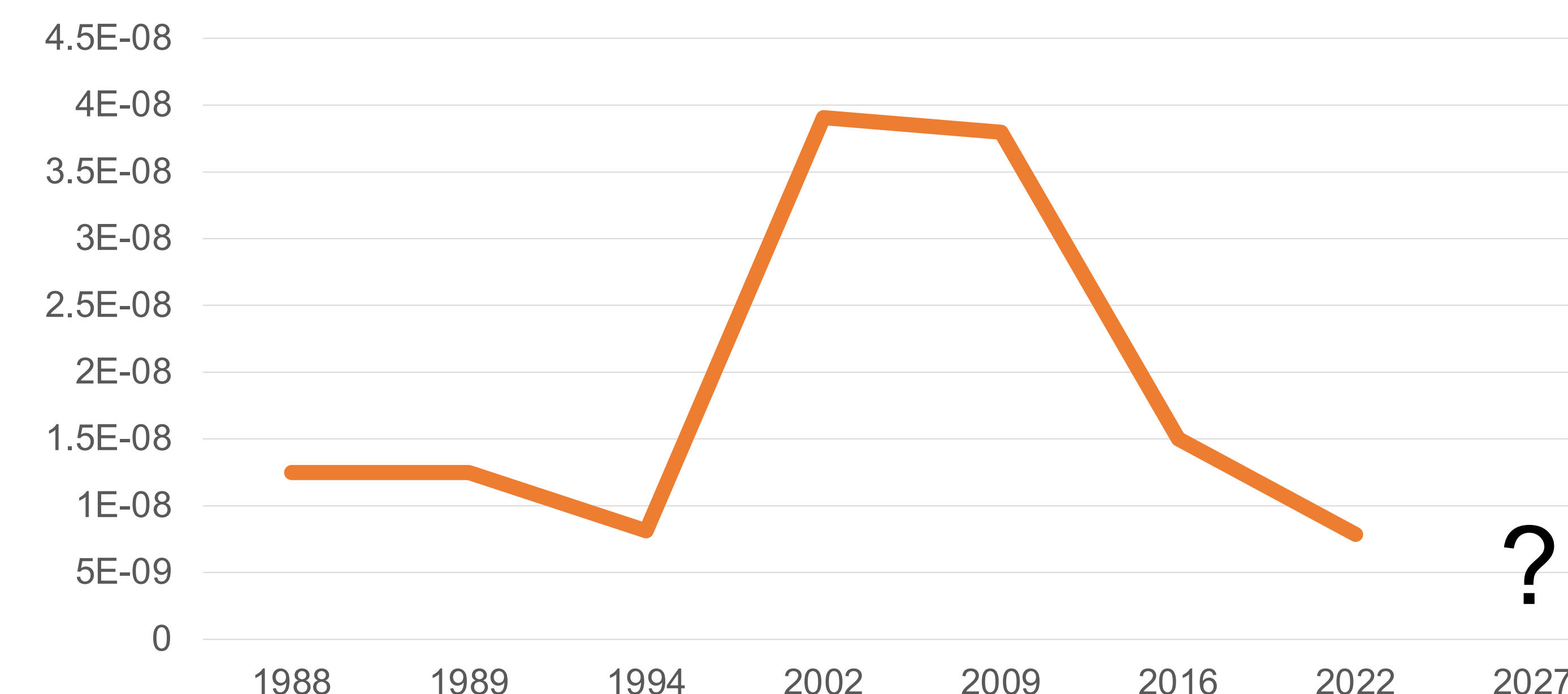
~5PB per compute project

Multiple of such projects work on a **single HPC infrastructure** at the same time:

~20-30PB just for high-res simulation data on one infrastructure

At the same time, **prices per unit of compute infrastructure** have stopped decreasing and are even increasing, leading to **less available storage per FLOP** given that the investment budget is constant

GB / FLOP Evolution at DKRZ



Every year marks the acquisition of a new HPC system at DKRZ

Source: DKRZ homepage, https://www.dkrz.de/en/systems/historie/computer-history-at-dkrz-1?set_language=en

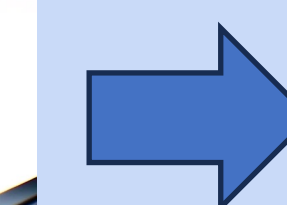
Data compression to the rescue?

Current ESM projects already utilize application-side lossless compression techniques, which help reduce storage space by roughly 40-50% (about factor 2 compression).

Lossy compression offers the potential for **higher compression rates (up to factor 1000)**, without access penalties for data retrieval.

We need **lossy compression**
But losing data accuracy is **scary**

To convince yourself that lossy compression is safe, you need to try it out yourself



An openly accessible Jupyter-based online laboratory for testing lossy compression techniques on ESM output datasets

The Online Laboratory for Climate Science and Meteorology

What's easier than opening a URL in your browser?

Serverless In-Browser Interactive Computing
jupyterlite
PYODIDE
WA

No setup, **no installation**, <1min to start

Reproducible and version-locked
Ensure your examples keep working

Supports many compiled scientific + especially Earth Science Python packages

Extra support for **accessing large datasets**

Ease of use: same code, same results

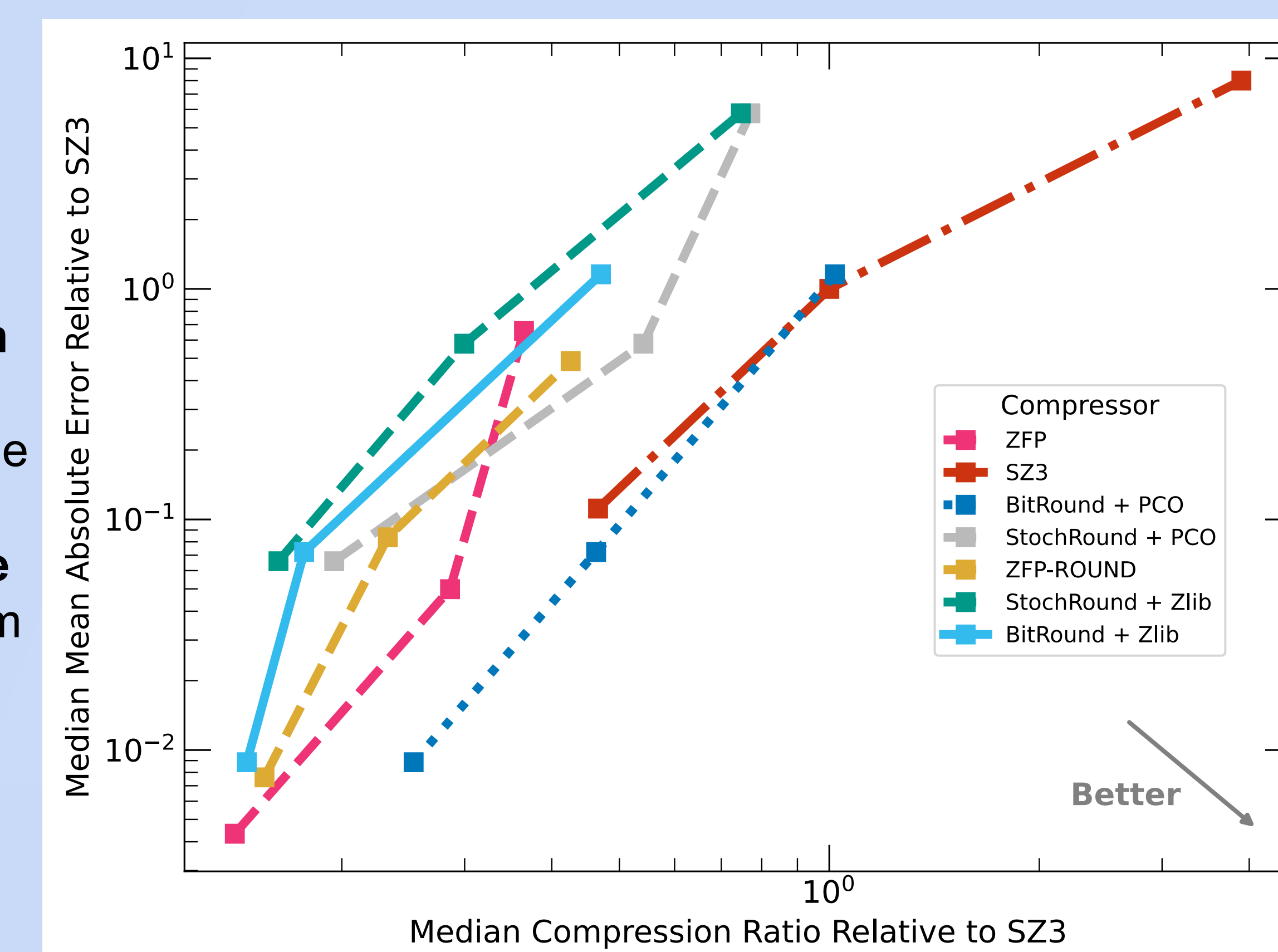
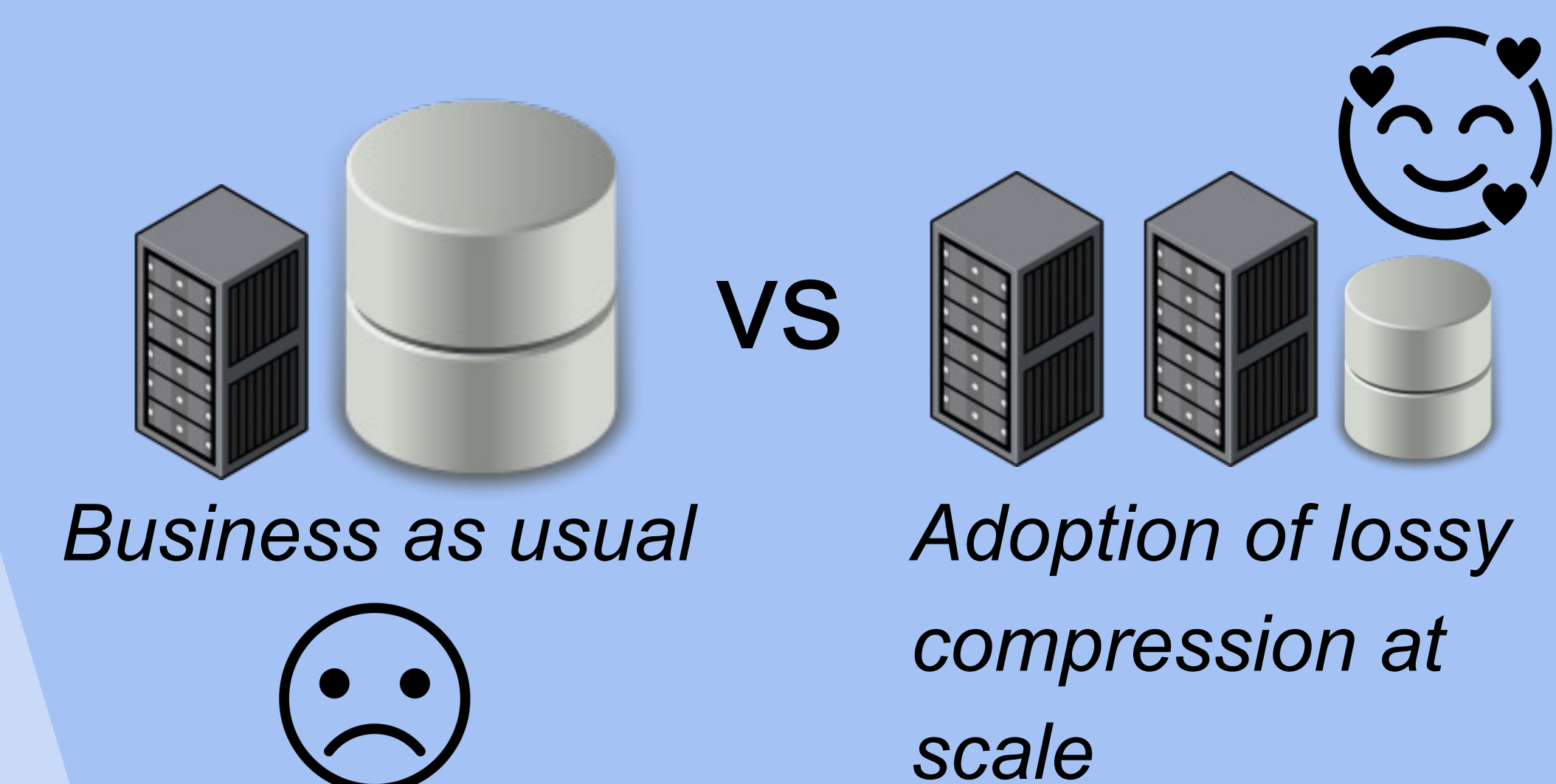
Internet icon by Freepik at https://www.flaticon.com/free-icon/internet_4861920

<https://compression.lab.climet.eu/>, currently offering compression testing for 7 lossy compressors

Impacts on HPC design decisions

Reduced data amounts **enable more efficient resource utilization** and allow for **smarter reinvestment of funds**

Given investment amount were constant (dEUR / dt = 0) **more powerful systems** enabling **exciting research** could be acquired



Sharing is encouraged